Chapter 2

Bayesian Hypothesis Testing for Single-Subject Designs

This chapter has been accepted for publication as:
de Vries, R. M., & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs, *Psychological Methods*, *18(2)*, 165 - 185. doi:10.1037/a0031037.[1]

Abstract

Researchers using single-subject designs are typically interested in score differences between intervention phases, such as differences in means or trends. If intervention effects are suspected in data, it is desirable to determine how much evidence the data show for an intervention effect. In Bayesian statistics, Bayes factors quantify the evidence in the data for competing hypotheses. We introduce new Bayes factor tests for single-subject data with two phases, taking serial dependency into account: a time-series extension of the Rouder et al.'s (2009) Jeffreys-Zellner-Siow (JZS) Bayes factor for mean differences, and a time-series Bayes factor for testing differences in intercepts and slopes. The models we describe are closely related to interrupted time-series models (McDowall et al., 1980)

2.1 Introduction

It is generally considered desirable in psychological research to collect data from as many participants as possible. However, in some situations, relevant questions may require careful observation of only a few subjects. Single-subject studies are useful in determining treatment effects when researchers or clinicians are interested in particular individuals, when they want to tailor interventions to individuals, when they want to carefully observe individuals separately rather than a group as a whole, or when limited resources do not permit group studies. In contrast to what the name implies, single-subject studies usually involve more than one subject, but the data are analyzed per subject and are not aggregated over a group of subjects. Typically, for each subject a sequence of baseline observations is available, together with a sequence of observations during and after one or more interventions. Morgan and Morgan (2009) give an extensive overview of design and data interpretation for single-subject studies.

Single-subject designs have, for example, been used to investigate the effect of a cognitive-behavioral intervention on depression after stroke (Rasquin et al., 2009) and the effect of nursing in implementing a behavior plan to reduce aggressive behavior (Bisconer et al., 2006). Beeson and Robey (2006) reported that of 620 studies concerning treatment approaches for aphasia and related disorders, 252 (41%) involved single-subject experimental studies. Kinugasa et al. (2004) plead for more use of single-subject designs in sport research to investigate intervention effects and predict performance for particular athletes.

Even when the interest is in group effects and data from a larger sample is available, additional focus on individuals can be informative. As Jacobson and Truax (1991), among others, have noted, overall effects observed in the group at large provide no information about the variability of treatment effects among the subjects. For instance, suppose that at post-measurement subjects from a treatment group have fewer symptoms than subjects from a control group, on average. Then it is not clear whether all subjects from the treatment group have improved, or whether some subjects from the treatment group have improved by a large amount while the remainder of the subjects have remained unchanged or have even deteriorated.

Analysis of single-subject data provides insight into how the subject has developed over time. For example, a continuous positive trend in the data shows a gradually increase in scores over time, and a sudden stable increase after an intervention suggests an intervention effect. Visual inspection of the data gives a first impression of possible intervention effects, but does not provide effect sizes or formal inferential evidence. Several effect size statistics and inferential techniques have been developed to augment visual inspection of single-subject data. For example, the percentage of non-overlapping data (Mastropieri and Scruggs, 1985) is an effect size based on the overlap of data points before and after an intervention, and several related measures have been developed like the percentage of data points exceeding the median of the baseline phase (Ma, 2006), percentage of all non-overlapping data (Parker et al., 2007), non-overlap of all pairs (Parker

and Vannest, 2009), and percentage of nonoverlapping corrected data (Manolov and Solanas, 2009). Other effect sizes for single-subject designs are the standardized difference of Shadish et al. (2008) and the effect size proposed by Maggin et al. (2011) based on generalized least squares regression, taking both the difference in intercepts and trends into account. Also some inferential techniques for single-subject data have been proposed, such as interrupted time series analysis (ITSA; McDowall et al., 1980) and permutation tests (Bulté and Onghena, 2008, 2009; Ferron and Foster-Johnson, 1998). However, of particular interest are inferential techniques that allow the evaluation of how much evidence exists in the data for or against the hypothesis that the intervention has had an effect. These techniques have not yet been developed for single-subject data.

In this paper, we develop methods for making inferences from single-subject data that we believe have advantages over current techniques, because they allow researchers to quantify the evidence for or against hypotheses about intercept and trend differences between intervention phases. The inferential technique we advocate is the Bayes factor, a part of the Bayesian statistical framework that has gained popularity in many fields in recent years. The Bayesian approach differs markedly in its assumptions from "classical" techniques that dominate the psychological literature; however, these different assumptions allow the development of techniques which address different questions, as we will see. For now, we take the Bayesian viewpoint for granted and leave comparison with other approaches for the Discussion. Our development represents, to our knowledge, the first application of Bayes factors to single-subject data.

In addition to theoretical development, we also report practical development in the form of easy-to-use software for the application of our technique. The techniques can be easily applied using the BayesSingleSub R package, which can be found at `http://cran.r-project.org/web/packages/BayesSingleSub`. A demonstration of how to use the R functions can be found in Section 1 of the online Supplement[2] to this article.

We introduce inference for single-subject data in the context of a simple example. In this introduction we present a simple Bayesian analysis of a mean difference, which we will subsequently extend to more complex and useful models for single-subject data. The introduction to Bayesian inference is included for readers who are unfamiliar with Bayesian methods. Readers familiar with Bayesian methods are invited to skip directly to "Bayes factors for single-subject data", where the extended models are presented. Finally, in the Discussion, we place our techniques in the context of current statistical practice, specifically null hypothesis significance testing.

[2]The online Supplement is included in this dissertation as Section 2.7

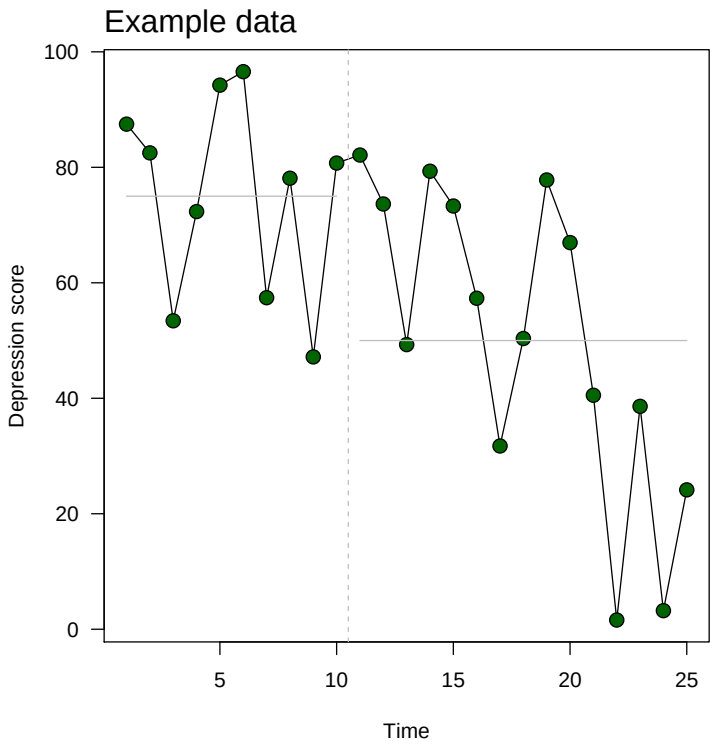

Figure 2.1: Example data; vertical line separates baseline and post-intervention phase, horizontal lines represent mean scores in each phase.

2.2 Introduction to inference for single-subject data

In single-subject research, a common question is whether a particular intervention is effective for a participant. We consider here research designs in which a participant is tested several times before an intervention (that is, at baseline). The participant is then tested several times during or after the intervention. To illustrate the measures of evidence we will discuss, we will make use of a fictitious example which has a similar form: a sequence of baseline observations followed by a sequence of observations after an intervention. A real-world example of such a design can be found in Rasquin et al. (2009).

Consider a hypothetical client who undergoes an intervention against depression. To monitor his progression he fills out a depression questionnaire ten times before and fifteen times after an intervention. A minimum score of 0 indicates no depression symptoms, and a maximum score of 100 indicates serious depression symptoms. Figure 2.1 shows the resulting data series. In the figure, the hori-

zontal lines show the mean scores at baseline and post-intervention. The average score at baseline is 75; after the intervention, the average depression score has dropped by 25 points to 50.

In this example our question of interest is whether the observed difference of 25 points is due to the effect of the intervention or whether it is simply due to random variation. We can regard these two possibilities as competing hypotheses. One hypothesis, the null hypothesis, states that there is no intervention effect. Another, which we call the alternative hypothesis, states that there is a true, nonzero intervention effect. Our task is to determine what the data tell us about the two competing hypotheses; or, equivalently, we wish to weigh the statistical evidence for one hypothesis against the evidence for the other.

In order to weigh statistical evidence, it is necessary to first describe a statistical model from which the data may have arisen. For instance, we might assume, as is often typical, that observations in each phase are independent of one another and distributed normally. For the sake of demonstration, we assume that the true standard deviation of the observations in each phase, which we denote as σ_ϵ, is equal to 24.5. We relax the assumptions of independence and known variance in later sections, where we extend this simple model to models for single-subject data. Our question of interest can be formulated in terms of the means of the two phases: we want to know whether the difference between the true means in Phases 1 and 2 is 0. In order to make the size of the difference easier to interpret, we divide the difference by the standard deviation to compute the standardized mean difference Cohen's d (Cohen, 1992):

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_\epsilon} = \frac{25}{24.5} \approx 1,$$

where $\bar{X}_1$ and $\bar{X}_2$ are the observed means of baseline and post-intervention conditions, respectively. Cohen's standardized measure of effect size d indicates that the mean difference is about 1 standard deviation. We can also formulate our research question and hypotheses in terms of the true, standardized effect size $\delta = (\mu_1 - \mu_2)/\sigma_\epsilon$, where μ_1 and μ_2 are the true means of the baseline and post-intervention phases, respectively. Under the null hypothesis that there is no intervention effect, $\delta = 0$; under the alternative hypothesis that there is an intervention effect, $\delta \neq 0$.

2.2.1 Likelihood ratios

In statistics, evidence from data is computed by means of the *likelihood ratio* (Hacking, 1965; Royall, 1997; Glover and Dixon, 2004). The likelihood ratio is a comparison of how likely the observed data are under the null and the alternative hypotheses. When the data are more likely under the null hypothesis than the alternative hypothesis, the data support the null; likewise, when the data are more likely under the alternative hypothesis than the null hypothesis, the data support the alternative. To make this concrete, suppose we are specifically interested in whether the true mean difference is 0 ($\delta = 0$; null hypothesis) or

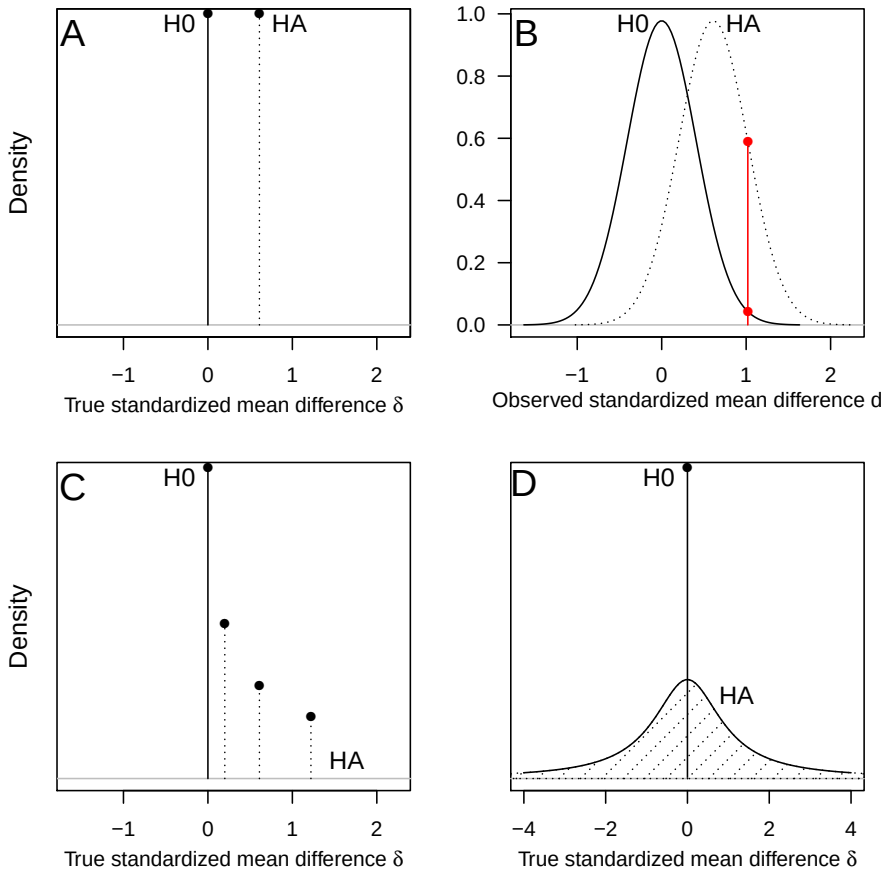

Figure 2.2: Illustrations of point null and point alternative hypothesis (A), likelihood ratio (B), composite alternative hypothesis consisting of three values (C), and continuous composite alternative hypothesis (D).

15 (δ=15/24.5=.61; an alternative hypothesis), as illustrated in Figure 2.2A. Because we have assumed the observations are normally distributed with known variance, we can compute the likelihood of the data under the two hypotheses using the probability density function for the normal distribution[3]. Figure 2.2B shows the likelihood of the data, summarized by Cohen's d, under each of the hypotheses. The two curves represent the sampling distribution of the observed standardized difference under the null (H_0) and the alternative (H_1) hypotheses. The likelihood of the data under the null hypothesis is .04, and the likelihood of the data under the specific alternative hypothesis (that $\delta = .61$) is .59. The likelihood ratio L in favor of the null hypothesis is

$$L = \frac{p(d \mid \delta_0, \sigma_\epsilon^2)}{p(d \mid \delta_1, \sigma_\epsilon^2)} = \frac{.04}{.59} = .07,$$

where $\delta_0 = 0$ and $\delta_1 = .61$ are the hypothesized values of δ under the null and alternative hypotheses, respectively. This is easy to see graphically in Figure 2.2B, where the density under the null hypothesis is .07 times that under the alternative.

When the likelihood ratio is less than 1, it is often useful to invert it to find the relative evidence for the alternative over the null. The likelihood ratio $L = .07$ indicates that the observed data are $1/.07 \approx 14$ times more likely under the alternative than under the null hypothesis. The evidence for $\delta = .61$ is 14 times stronger than the evidence for $\delta = 0$. To interpret the likelihood ratio, we note that a likelihood ratio of 1 means that the data are equally likely under either hypothesis, and thus the data favor neither hypothesis. The more the likelihood ratio deviates from 1, the stronger is the evidence in the data for the null or alternative hypothesis. The amount of evidence the data contain is affected by several factors, including the number of data points and the variation of the data points within a phase. The larger the number of data points and the less variation within the phases, the more the data allow us to differentiate between the hypotheses.

In order to calculate the likelihood ratio we specified the alternative hypothesis specifically as $\delta = .61$. In practice, however, the alternative hypothesis is typically not a single value. It is therefore desirable to specify a composite alternative hypothesis consisting of several plausible, nonzero effect sizes. For example, we could specify an alternative hypothesis that consisted of three possible true differences $\mu_1 - \mu_2$: 5 ($\delta = .20$), 15 ($\delta = .61$), or 30 ($\delta = 1.22$). In this case, if the alternative hypothesis is true, we are only willing to entertain the possibility that δ is one of these three values. One way to compute the likelihood under this new composite alternative is to compute the likelihood of the data under each specific value, and then compute a weighted average likelihood. These weights will depend on relative plausibility of each hypothesized value of δ. For instance, a reasonable set of weights for the three hypothesized values of δ under the alternative hypothesis might be .5, .3, and .2, respectively, reflecting the expectation

[3]The probability density function for the normal distribution is $p(y) = \exp\{-(y - \mu)^2/(2\sigma^2)\}/\sqrt{2\pi\sigma^2}$.

that smaller effect sizes are more plausible than larger effect sizes. Note that these weighting values sum to 1, and thus may be thought of as probabilities. This quantification of uncertainty about the value of parameters as probability reflects a unique property of Bayesian statistics (Jeffreys, 1961; Jaynes, 1986; Wagenmakers et al., 2008). The weighted effect sizes together would form our composite alternative hypothesis, which instead of consisting of only one value, consists of three values each with a different plausibility. The null and composite alternative hypotheses are illustrated in Figure 2.2C.

Having defined the alternative hypothesis, we can calculate the weighted average of the likelihoods over the values of the alternative. In our example the likelihood of the data for $\delta = .20$ is .13, for $\delta = .61$ the likelihood is .59, and for $\delta = 1.22$ the likelihood is .87. The weighted average likelihood is thus

$$\sum_{i=1}^{3} p(d \mid \delta_i, \sigma_\epsilon^2) p(\delta_i) = (.5 \times .13) + (.3 \times .59) + (.2 \times .87) \approx .42,$$

where δ_i is the ith possible value of δ under the alternative, and $p(\delta_i)$ a function giving the weights for the corresponding δ_i values. This weighted average of likelihoods represents the likelihood of the data under the composite alternative hypothesis. In Bayesian statistics, this weighted average likelihood has a special name: the *marginal likelihood*. In our example, the computation of the marginal likelihood of the data is justified by interpreting the weights as probabilities.

Having computed the marginal likelihood, we can compare the likelihood of the data under the null and composite alternative hypothesis and calculate the marginal likelihood ratio. Because the null hypothesis is only a single value, $\delta = 0$, its likelihood remains the same. Hence the marginal likelihood ratio for the null versus the composite alternative hypothesis equals $.04/.42 \approx .1$. The observed data are about .1 times as likely to occur under the null than under the composite alternative. Stated otherwise, the data are $1/.1 = 10$ times more likely under the composite alternative than under the null. As before, this is strong evidence for the alternative hypothesis over the null hypothesis.

The composite alternative hypothesis we considered with three possible values is more flexible than the alternative hypothesis containing only one value for δ. However, it still forces the researcher to choose a small set of effect sizes and corresponding weights. Luckily, because in Bayesian statistics the weights are interpreted as probabilities, it is straightforward to extend the composite hypothesis to include all real numbers: instead of defining a discrete weighting distribution for δ, we use a continuous weighting function. An example is shown in Figure 2.2D. The bell-shaped probability distribution shown in the figure is one possible weighting distribution for the unknown δ parameter. Values of δ near 0 are considered more likely than those farther away from 0, quantifying the expectation that large effect sizes are unlikely. Different expectations with regard to plausible effect sizes would require a different weighting distribution, matching these expectations. The expectations determining the weighting distribution could be based on all kinds of considerations, like previous research,

expert knowledge, scale boundaries, statistical considerations, etc. In Bayesian statistics, the weighting distribution is called the *prior distribution*, because it quantifies the *a priori* likelihood of different values of δ being true.

With a continuous distribution as our weighting function, instead of summing over a small set of possible values of δ, we must integrate the likelihood over all real numbers. For readers without a calculus background, this can be thought of as directly analogous to the summing we demonstrated with the composite hypothesis containing only three values. In order to keep our exposition as non-technical as possible, we refer readers interested in more details to the excellent introduction to Bayesian statistics by Lee (Lee, 2004). After integration, we again obtain the marginal likelihood for the alternative hypothesis; for the weighting distribution shown in Figure 2.2D, the marginal likelihood is .17. This marginal likelihood is smaller than the marginal likelihood for the weighting distribution of Figure 2.2C, because many of the values considered plausible *a priori*, predict that the observed data are unlikely. Including these implausible values in the weighting distribution attenuates the weighted average likelihood. Again, we form a ratio of marginal likelihoods to compare the null to the alternative hypothesis, which is $.04/.17 \approx .24$. In Bayesian statistics, this ratio, called the *Bayes factor*, quantifies the extent to which the data support the null hypothesis over the alternative hypothesis. Throughout this paper, we denote the Bayes factor by B. In this case, the observed data support the hypothesis that $\delta \neq 0$ by a factor of about 4, because $B \approx 1/4$. This value, of course, must always be interpreted with the chosen weighting distribution in mind; we provide more discussion of this point later.

As a statistic, the Bayes factor, like the simpler likelihood ratio, is straightforward to interpret. The data support the null hypothesis when the Bayes factor is larger than 1 and support the alternative hypothesis when the Bayes factor is smaller than 1. The more the Bayes factor deviates from 1, the stronger the evidence for the null or alternative hypothesis. The Bayes factor is also the extent to which a rational person should adjust their beliefs, expressed as relative odds, in favor of the null hypothesis in response to the data. In the example above, if a researcher believed, prior to observing the data with Cohen's $d \approx 1$, that the null and (continuous) alternative hypotheses were equally likely, then that researcher should, after observing the data, hold 4 to 1 odds in favor of the alternative hypothesis. In Bayesian statistics, beliefs (again, expressed as relative odds) after observing the data are called *posterior* odds. However, we stress that the Bayes factors themselves are not odds, nor is positing prior and posterior odds required for interpreting the Bayes factor. If prior odds are stipulated, however, then the Bayes factor can be used to compute the posterior odds. See Appendix A for a discussion of posterior odds.

Researchers applying the Bayes factor may wonder when a Bayes factor implies weak or strong evidence for an hypothesis, that is, when a Bayes factor is small or large. Researchers familiar with p values often use conventional cutoffs to decide how to interpret a p value, such as that $p < .05$ indicates "significance". This is useful for p values, because p values cannot be interpreted as evidence by

themselves. Bayes factors, however, have a clear interpretation without recourse to verbal labels or cutoffs: a Bayes factor is the relative likelihood of the data under two different hypotheses, which then has the straightforward interpretation as being the degree to which a rational person will shift their beliefs on seeing the data. Whether a Bayes factor is large or small depends on the prior beliefs it is modifying. A Bayes factor of 1/20 is small when it is modifying prior odds of 1000000 (e.g., regarding clairvoyance), but large when it is modifying prior odds of, say, 2. Thus, whether a Bayes factor is small or large depends not only on its value but also on the context in which it is applied.

In the development above, we made certain simplifying assumptions to make introducing the Bayes factor easier; for instance, the variance of the observations was known, and the observations were independent. For the Bayes factor to be useful in single-subject research, these assumptions must be relaxed. Also, we only focused on mean change while single-subject researchers are often interested in other data patterns as well, like trend differences. In the next section, we review the Bayes factor t test of Rouder et al. (2009), which relaxes the assumption of known variance. For single-subject research, we must extend this Bayes factor to account for dependencies between time points (Busk and Marascuilo, 1988; Sharpley and Alavosius, 1988; Matyas and Greenwood, 1997) and for trend and intercept differences in the data. These extensions yield Bayes factors that are broadly applicable to single-subject designs.

2.3 Bayes factors for single-subject data

The Bayes factor for δ we develop is an extension of Rouder et al.'s (2009) Bayesian t test; we therefore first present the Rouder et al.'s Bayes factor. We then extend this Bayes factor for single-subject data and ultimately present the Bayes factors for trend and intercept differences.

2.3.1 Rouder et al.'s Bayes factor t test

We consider again the design described in the previous section, while making it more general. Suppose a participant provides n_1 baseline measurements and n_2 post-intervention measurements, for a total of $N = n_1 + n_2$ observations. We denote these observations as y_i $(i = 1, \ldots, N)$. As before, we assume that these measurements are normally distributed with a common variance σ_ϵ^2, and that an observation's true mean depends on the phase in which it was observed. We express this in conventional regression notation with a grand mean μ_0 and effect size δ:

$$\begin{aligned} y_i &= \mu_0 + \sigma_\epsilon \delta x_i + \epsilon_i, \\ \epsilon_i &\overset{\text{iid}}{\sim} \text{Normal}(0, \sigma_\epsilon^2), \end{aligned}$$

where the x_i are dummy codes. The first n_1 values of x are equal to -.5, and the final n_2 values of x are equal to .5. This coding, along with the multiplication

of δx_i by σ_ϵ, allows the parameter δ to be interpreted as the true standardized difference between the means of the two phases.

As in our previous example, under the null hypothesis $\delta = 0$, and under the alternative hypothesis $\delta \neq 0$. As before, we must choose a prior distribution for δ to serve as a function to weight the plausibility of different values of δ if the alternative is true. Rouder et al., following Jeffreys (1961) and Zellner and Siow (1980), used a t distribution with one degree of freedom. This distribution, also called the Cauchy distribution[4], is shown in Figure 2.3 (solid line). It is also the weighting distribution of Figure 2.2D. The reasons for the choice of the Cauchy distribution over other plausible prior distributions are technical and we will not cover them here (see Zellner and Siow, 1980, for details). Note, however, that the distribution generally comports with expectations about standardized effect sizes: the most likely values are around 0, and the plausibility of values drops rapidly as $|\delta|$ gets larger. In Bayesian statistics, the fact that δ has a Cauchy prior distribution is denoted:

$$\delta \sim \text{Cauchy}(r),$$

where r is a scaling factor. This notation, in which an unknown parameter has a probability distribution, makes explicit the notion in Bayesian statistics that the uncertainty in parameters can be expressed using probability. The scaling factor r allows the adjustment of the weighting distribution for different areas of study, across which plausible effects may vary. It can be interpreted as half the inter quartile range (IQR) of the Cauchy distribution, which is the distance between the first and third quartile of the distribution, represented by the red dots in Figure 2.3. In other words, two times the scaling factor r equals the IQR.

Rouder et al. recommend to use $r = 1$ by default, corresponding to $IQR = 2 \times 1 = 2$. From the figure, it may appear that $r = 1$ puts unrealistically large weight on large effect sizes, effect sizes that are usually not encountered in group studies. However, in single-subject studies effect sizes tend to be larger (Beeson and Robey, 2006; Parker et al., 2005, 2007; Parker and Vannest, 2009), which we believe makes the $r = 1$ default more reasonable for single subject studies. Still, plausible effect sizes may vary from study to study, and the r scale can be adjusted accordingly. Increasing r to 2 increases IQR to $2 \times 2 = 4$, resulting in a prior that puts more weight on larger effect sizes, as shown by the dotted line in Figure 2.3. Similarly, decreasing r to .5 decreases IQR to $2 \times .5 = 1$, resulting in a prior that puts more weight on smaller effect sizes, as shown by the dashed line of Figure 2.3.

Readers familiar with Bayesian statistics may have encountered so-called "flat," or "noninformative" priors, used in Bayesian parameter estimation to minimize the influence of the prior distribution on the parameter estimates. These priors are often used to quantify the idea that we have no *a priori* expectations regarding plausible values of a parameter. Flattening out the prior by increasing

[4]The Cauchy distribution has density function $p(\delta; r) = \left((1 + \left[\frac{\delta}{r}\right]^2)r\pi\right)^{-1}$.

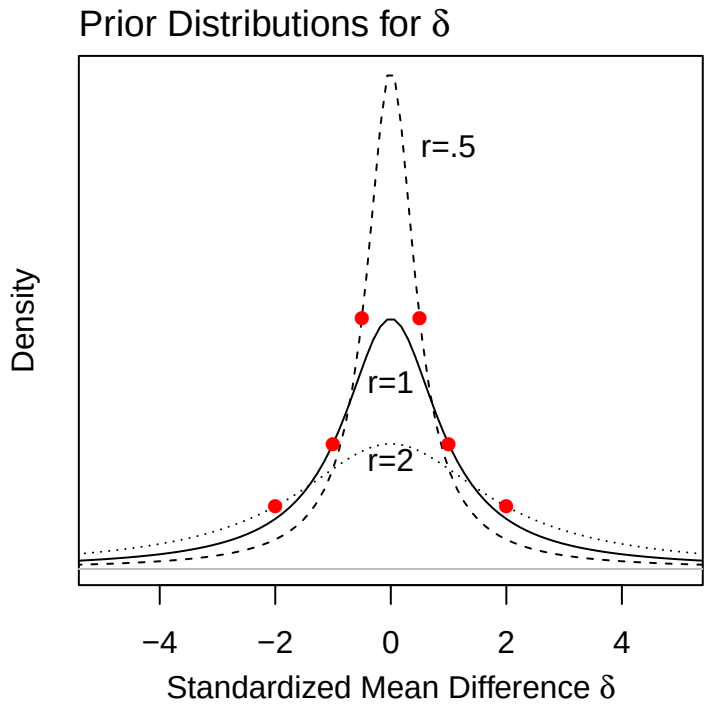

Figure 2.3: Cauchy distributions with scaling factors $r = .5$ (dashed line), $r = 1$ (solid line), and $r = 2$ (dotted line). Horizontal distances between red dots represent inter quartile ranges, equaling twice the scaling factor r.

r may seem desirable at first, by putting equal weight on all effect sizes. However, in hypothesis testing with Bayes factors, using a flat prior for a parameter under one hypothesis but not under the competing hypothesis is a mistake; a flat prior on δ under the alternative hypothesis puts too much weight on unrealistically large effect sizes, which makes the marginal likelihoods arbitrary small. This in turn makes the Bayes factor favor the null hypothesis. This is, of course, not surprising; if our alternative hypothesis entertains the idea that a standardized effect size of $\delta = 1,000,000$ is as likely as $\delta = 0$, it should be rejected. We thus advocate using $r = 1$ by default, unless some justification can be given for another value of r.

Although we have defined a prior distribution for δ under the alternative hypothesis, we cannot yet compute a Bayes factor. In the previous example, we assumed that the within-phase variance σ_ϵ^2 was known; in order to make the analysis useful in practice, we must drop this assumption. To compute the Bayes factor, we must compare the marginal likelihood under two hypotheses: the null hypothesis and the alternative hypothesis. Although the null hypothesis specifies

that $\delta = 0$, neither hypothesis gives any constraint on possible values of σ_ϵ^2. In both models, σ_ϵ^2 is a nuisance parameter. In Bayesian statistics, nuisance parameters are treated in the same way as parameters of interest: prior distributions are stipulated, which can then be averaged out through integration. For μ_0 and σ_ϵ^2, Rouder et al. used a joint prior suggested by Jeffreys (1946):

$$p(\mu_0, \sigma_\epsilon^2) \propto \frac{1}{\sigma_\epsilon^2},$$

where $\propto$ means "proportional to". This prior is a standard prior in Bayesian statistics, due to a special property: namely, that it is scale-invariant[5]. This scale-invariance produces Bayes factors which do not depend on the units of the dependent variable. Thus, linear transformation of the dependent variable will not affect the Bayes factor.

With priors defined on all parameters, it is possible to define the Bayes factor statistic. We use integration to average out the unknown parameters μ_0 and σ_ϵ^2 under the null hypothesis and μ_0, σ_ϵ^2, and δ under the alternative hypothesis, and construct the ratio of the two marginal likelihoods:

$$B = \frac{\int\int p(\mathbf{y} \mid \delta = 0, \mu_0, \sigma_\epsilon^2) p(\mu_0)\, p(\sigma_\epsilon^2)\, d\mu_0\, d\sigma_\epsilon^2}{\int\int\int p(\mathbf{y} \mid \delta, \mu_0, \sigma_\epsilon^2) p(\delta)\, p(\mu_0)\, p(\sigma_\epsilon^2)\, d\delta\, d\mu_0\, d\sigma_\epsilon^2},$$

where $\mathbf{y}$ represents the entire vector of data. As mentioned in the previous section, these integrals are analogous to the weighted average we computed in the previous section, but over the entire plausible range of the prior distribution. Rouder et al. called this Bayes factor the Jeffreys-Zellner-Siow (JZS) Bayes factor, to reflect its origin in the work of these three statisticians. We abbreviate this Bayes factor B_{jzs} to differentiate it from other Bayes factors we define subsequently.

After simplification, B_{jzs} can be conveniently written as a function of only the traditional two-sample t statistic and the effective sample size N_0:

$$B_{jzs} = \frac{\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}}{\int_0^\infty (1+N_0 g)^{-1/2} \left(1 + \frac{t^2}{(1+N_0 g)\nu}\right)^{-(\nu+1)/2} (2\pi)^{-1/2} g^{-3/2} e^{-1/(2g)} dg}, \quad (2.1)$$

where the degrees of freedom $\nu = n_1 + n_2 - 2$ and effective sample size N_0 is

$$N_0 = \frac{n_1 n_2}{n_1 + n_2}.$$

The parameter g is introduced for convenience of integration (see further details in Section 2 of the online Supplement). Although Eq. 2.1 may appear

[5]Observant readers will notice that this is a "noninformative" prior of the type we warned about in a previous paragraph. The use of flat priors is acceptable in Bayes factors when the noninformative priors are placed on parameters that are not targets of inference.

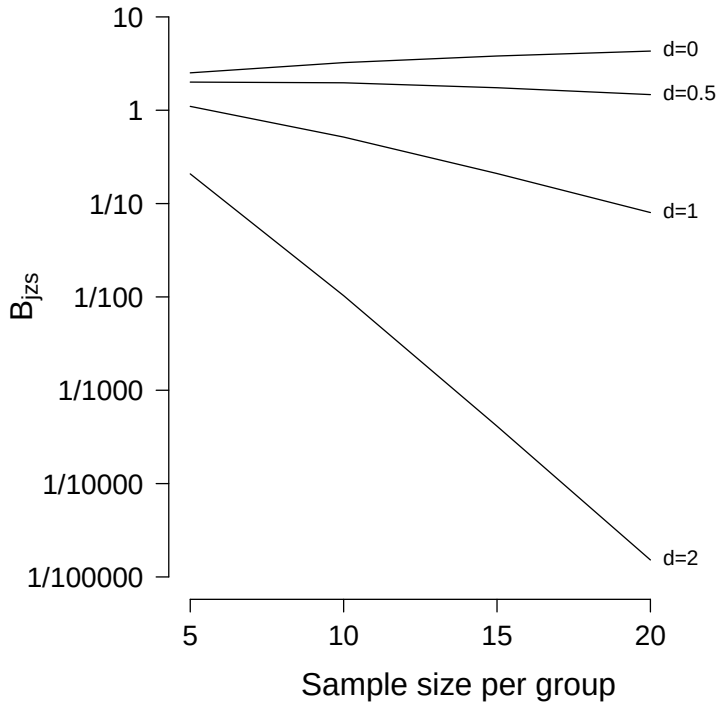

Figure 2.4: B_{jzs} as a function of observed Cohen's d for different sample sizes per group, with $r = 1$.

complicated, it is straightforward to evaluate using software that can perform one-dimensional integrals, such as Microsoft Excel or R (R Development Core Team, 2009). Rouder et al. provide an easy-to-use web applet to compute the Bayes factor at `http://pcl.missouri.edu/bayesfactor`. The user need only provide a t statistic, which is obtainable from any common statistical program, and the sample sizes n_1 and n_2. The Bayes factor provided by the website can then be used to evaluate the evidence for the null hypothesis that $\delta = 0$ relative to the alternative hypothesis that $\delta \neq 0$, with the Cauchy prior distribution on δ as the alternative. It should be noted that the fact that B_{jzs} is a function of the t statistic and sample size does not mean that it is essentially the same as a t statistic or p value. The t statistic and sample size summarize information in the data, and the classical t test and B_{jzs} use this information in different ways, resulting in different numbers with different meanings.

Figure 2.4 shows how the B_{jzs} changes as a function of observed Cohen's d, for different sample sizes commonly found in single-subject research (Parker and Brossart, 2003; Parker and Hagan-Burke, 2007; Parker and Vannest, 2009). The figure shows that when observed Cohen's d is 0, B_{jzs} slowly increases from 1 as the

sample size increases. That is, with an observed Cohen's d of 0, the B_{jzs} provides more support for the null hypothesis when the information in the data increases. For a Cohen's d of .5 these sample sizes are too small to obtain much evidence for the alternative: The B_{jzs} moves from 2 to 1.5 for this effect size. With so little information in the data there is not enough evidence for the alternative and, if anything, the null hypothesis is slightly supported because it provides a more parsimonious explanation of the data than the alternative hypothesis. With larger sample sizes, however, the Bayes factor for $d = .5$ would eventually favor the alternative hypothesis. For larger effect sizes the B_{jzs} decreases from 1 as sample size increases. The more information the data contain, the more evidence for the alternative hypothesis.

For the example data of Figure 2.1 the JZS Bayes factor equals .27. This means that it favors the alternative hypothesis that δ differs from zero: the data are $1/.27 = 3.7$ times more likely under the alternative than under the null. Note that the JZS Bayes factor of .27 is closer to 1 than the Bayes factor we calculated based on the continuous weighting distribution for δ of Figure 2.2D. This is not surprising, as even though for both Bayes factors the prior distribution for δ is a Cauchy distribution, the JZS Bayes factor does not assume that the true within-phase variance σ_ϵ^2 is known. Rather, it takes into account the extra uncertainty due to the unknown σ_ϵ^2, yielding a slightly different Bayes factor.

Although the JZS Bayes factor provides a powerful tool for evaluating the relative evidence in the data for the null versus the alternative hypothesis, it makes the strong assumption that observations are independent. Although this may be a useful assumption in some research, it is not a realistic one in single-subject research. In single-subject data, the independence assumption is problematic, because the data are measurements from a single-subject across several time points. Measurements at two adjacent time points are likely to be more similar than two measurements at nonadjacent time points, a type of dependency called *positive serial autocorrelation* (Fox, 2008, chap. 16). Due to the shared information across time-points, the effective sample size in the data is lower than the actual sample size. The JZS Bayes factor is therefore an over-estimation of the amount of evidence in the data. In the next section, we extend the JZS Bayes factor to account for serial dependencies across time-points, and thus make it more appropriate for the analysis of single-subject data.

2.3.2 Rouder et al.'s Bayes factor t test extended for time-series data

To make the JZS Bayes factor applicable for single-subject data, we extended the underlying model such that it accounts for serial dependency. More specifically, we extended the JZS model such that the errors come from a lag 1 auto-regressive (AR(1)) process and therefore we call it the JZS+AR model. In a lag 1 auto-regressive process, the error in one observation depends on both the previous error and on an independent and randomly drawn value, which we will denote z. The latter part is also called the random "shock." The level of serial dependency

can be increased by increasing the contribution of the previous error relative to the contribution from the random shock. The AR(1) model is a special case of the Auto-Regressive Integrated Moving Average (ARIMA) model (Fox, 2008, chap. 16), which Parker et al. (2005) suggested for use in inference for single-subject data. We choose the specific AR(1) process because it is the most commonly used process to model serial dependencies in the social sciences (Fox, 2008, chap. 16); however, our development is sufficiently general that other processes can be accommodated by changing the specified correlation matrix.

Formally, the JZS+AR model is the same as the JZS model, except for the addition of a parameter which controls the amount of correlation between successive time points. We add this parameter to both the null and the alternative hypothesis, because even in the absence of an intervention effect under the null model, there is reason to believe that measurements across time points will be serially dependent. As before, the data are a function of a grand mean parameter μ_0, an intervention effect δ, and random error ϵ:

$$y_i \quad = \quad \mu_0 + \sigma_z \delta x_i + \epsilon_i,$$

where $\delta = 0$ under the null hypothesis. σ_z is the standard deviation of the normally distributed random shocks z_i, and is similar to the σ_ϵ in the JZS model. The only difference is that the σ_ϵ in the JZS model fully represents the standard deviation of the errors, while the σ_z in the JZS+AR model is only a part of it. This is because in the JZS+AR model, we add covariances between the errors ϵ_i:

$$\epsilon \quad \sim \quad \text{Multivariate Normal}_N(\mathbf{0}, \sigma_z^2 \mathbf{\Psi}),$$

where $\mathbf{0}$ is a vector of zeros and $\mathbf{\Psi}$ is a correlation matrix. The particular form of the correlation matrix $\mathbf{\Psi}$ we choose is determined by the AR(1) process:

$$\mathbf{\Psi} = \frac{\rho^{\mathbf{D}}}{1 - \rho^2},$$

where ρ is the true correlation between successive time points, also called the lag 1 autocorrelation. The matrix $\mathbf{D}$ is a matrix of ordinal distances between time points:

$$\mathbf{D} = \begin{pmatrix} 0 & 1 & 2 & \cdots & N-1 \\ 1 & 0 & 1 & \cdots & N-2 \\ 2 & 1 & 0 & \cdots & N-3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ N-1 & N-2 & N-3 & \cdots & 0 \end{pmatrix},$$

and $\rho^{\mathbf{D}}$ represents a matrix where the ijth element is $\rho^{D_{ij}}$. This matrix of distances $\mathbf{D}$ allows the correlation between any two time points to decrease as a function of their distance from one another.

The prior distributions for all parameters of the JZS+AR model are the same as those for corresponding parameters of the JZS model. However, the JZS+AR model has an additional parameter ρ for the lag 1 autocorrelation. Because it is unlikely that we will ever know the true value of the lag 1 autocorrelation, we place a prior distribution on ρ that captures reasonable expectations for what the true value might be in single-subject data. In practice, auto-correlation in single-subject data is found to be mainly positive and reasonably low, smaller than about .3 (Fox, 2008, chap. 16; Matyas and Greenwood, 1997; Parker et al., 2005); we therefore choose a prior distribution which bars the possibility of negative values, and weights lower values of ρ more than higher values. The particular prior distribution for ρ that we advocate is shown in Figure 2.5 (solid line). This distribution is called a beta(a,b) distribution[6] (Casella and Berger, 2002); in particular, it is a beta distribution with $a = 1$ and $b = 5$. This prior distribution on ρ reflects the expectation that ρ is likely to be low, but might take on values as high as .4 or .5. Setting a to 1 ensures that the density of the beta distribution always decreases as ρ moves from 0 to 1. Setting b to 5 ensures that large ρ values are considered unlikely but not practically impossible. Increasing b would result in a distribution putting less weight on larger values, while decreasing b would result in a distribution putting more weight on larger values. It should be noted that the specific choice of the beta(1,5) prior distribution on ρ is not fundamental to the Bayes factor; researchers with different expectations for ρ can choose different prior distributions which are reasonable for their field of research.

With the prior distribution for ρ defined, it is possible to compute the Bayes factor for the null hypothesis $\delta = 0$ versus the alternative hypothesis that $\delta \neq 0$, with serial dependencies included. We abbreviate this Bayes factor B_{ar} to distinguish it from the JZS Bayes factor. Although the formula for the JZS Bayes factor was a function of t and the sample sizes alone (Eq. 2.1), the JZS+AR Bayes factor has no corresponding simple formula. The reason is that the JZS+AR Bayes factor takes into account the pattern of residuals in the data. The integration required to compute the JZS+AR Bayes factor is thus considerably more complicated. Details of how the Bayes factor is computed can be found in Section 2 of the Supplement to this article. The BayesSingleSub R package contains easy-to-use R functions for computing the Bayes factor. Because some readers may be new to using R, we provide a document demonstrating the use of our software in Section 1 of the online Supplement.

To demonstrate the properties of the JZS+AR Bayes factor, and to compare it to the JZS Bayes factor, we simulated data with several levels of effect size, sample size, and positive auto-correlation, and analyzed these data with both Bayes factor statistics. The details of the simulations can be found in Appendix B. To examine how the JZS+AR Bayes factor uses the information in the prior distribution for ρ, we calculated B_{ar} using three different prior distributions for ρ: a flat prior (beta(1,1), Figure 2.5, dashed line), the beta(1,5) distribution, and a prior that puts more weight on auto-correlations closer to zero (beta(1,15),

[6]The beta distribution has distribution function $p(\rho) = \Gamma(a+b)/(\Gamma(a)\Gamma(b))\rho^{a-1}(1-\rho)^{b-1}$, where Γ is the gamma function (Abramowitz and Stegun, 1965).

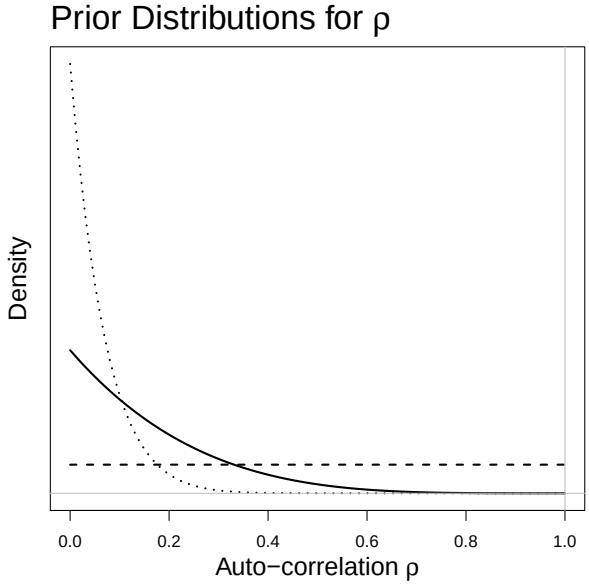

Figure 2.5: Comparison of beta prior distributions for the auto-correlation ρ; $a = 1$, $b = 1$ (dashed line), $b = 5$ (solid line), and $b = 15$ (dotted line).

Figure 2.5, dotted line).

Figure 2.6 shows B_{jzs} (dashed lines) and B_{ar} (solid lines and blue dots) as a function of observed absolute Cohen's d for different sample sizes per phase N and different prior distributions for the auto-correlation ρ. The solid lines are nonparametric regression lines (LOWESS; Cleveland, 1981). The dots show the variation in the B_{ar} due to the patterns in the data. Note that the B_{jzs} is not sensitive to auto-correlation and therefore all B_{jzs} values are the same given an effect size and sample size.

The figure shows how the JZS+AR Bayes factor penalizes the evidence in the presence of auto-correlation. The Bayes factors B_{ar}, which account for auto-correlation, are attenuated relative to the JZS Bayes factor B_{jzs}, which does not take auto-correlation into account. This attenuation is due to the fact that the effective sample size is decreased by the presence of auto-correlation; the amount of evidence in the data for either hypothesis is substantially less than the actual sample size might imply. The JZS Bayes factor therefore overestimates the evidence in the data. Figure 2.6 also shows how B_{ar} is affected by the choice of the prior distribution for the auto-correlation. When the prior distribution puts increasingly more weight on smaller auto-correlations ($b = 15$), B_{ar} assumes a lower level of auto-correlation on average, and thus provides less penalization of the evidence for an intervention effect. Hence B_{ar} is closer on average to the B_{jzs} when the prior distribution on ρ is concentrated closer to 0. As the

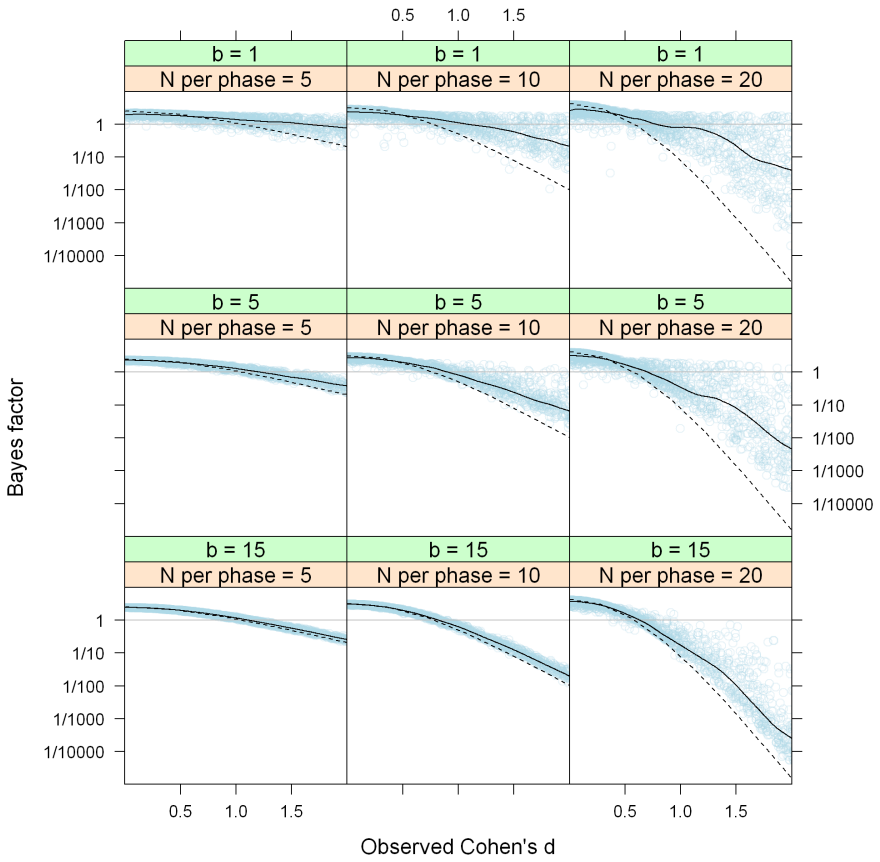

Figure 2.6: B_{jzs} (dashed line) and B_{ar} (solid lines and blue dots) as a function of observed Cohen's d for different sample sizes and prior distributions for the auto-correlation; solid lines are nonparametric regression lines.

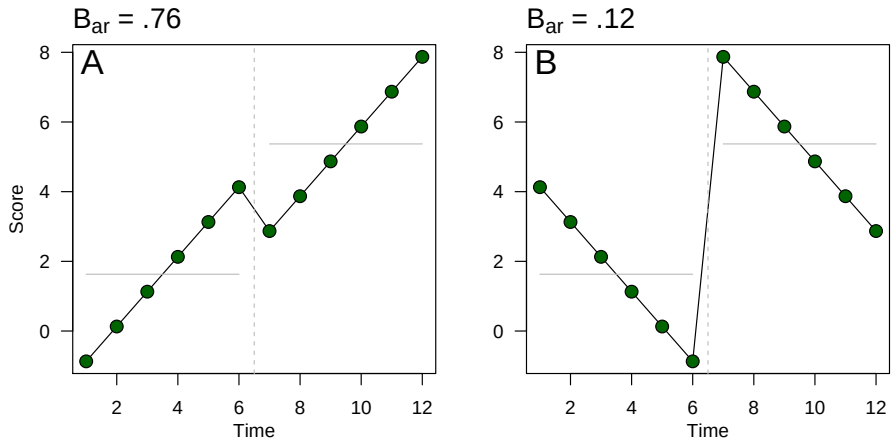

Figure 2.7: Two data sequences consisting of the same data per phase but in reversed order; AR Bayes factors for the left and right sequences are .76 ($1/.76 = 1.3$) and .12 ($1/.12 = 8.3$), respectively.

prior parameter $b \to \infty$, B_{ar} will penalize the evidence less and less, eventually converging to B_{jzs}.

It is clear from Figure 2.6 that there is substantial variation in the JZS+AR Bayes factor, even for the same observed Cohen's d and sample size. For example, for observed Cohen's d of about 2, for $b = 1$ and $n_1 = n_2 = 20$ (Figure 2.6, top right), the JZS+AR Bayes factors range between about 3 in favor of the null hypothesis of no intervention effect, and 1000 in favor of an intervention effect. It is reasonable to ask what is driving this variation of several orders of magnitude in the evidence for an effect. In order to answer this question, one can examine sequences of data points with the same observed Cohen's d and sample size, but different orderings of points, as in Figure 2.7. The figure shows two data sequences, each with six data points at baseline and six data points after the intervention. The two data sequences consist of exactly the same values per phase, and thus have an equal Cohen's d. The only difference between the two sequences is that the data within a phase are in reversed order. Still, the B_{ar} is different for the two data sequences: for the sequence in Panel A, $B_{ar} = .76$, thus favoring the alternative by a factor of 1.3. This represents equivocal evidence, as neither model is favored very strongly. For the sequence in Panel B, however, the evidence for the alternative is much stronger: $B_{ar} = .12$, favoring the alternative by a factor of 8.3.

To explain this behavior, we focus on the specific data patterns. For both patterns there are two plausible explanations, which the Bayes factor balances against one another. The first explanation is that there is positive auto-correlation and an intervention effect (true mean difference between the two phases; alternative hypothesis). The positive auto-correlation would explain the upward or

downward trends while the intervention effect would explain the overall shift of the data in the second phase. The second explanation is that there is only positive auto-correlation and no intervention effect (null hypothesis). The positive auto-correlation would explain the upward or downward trends and the negatively related data points (large values followed by small values and vice versa) would reflect random noise. The first explanation implies an intervention effect, but the second explanation does not.

Considering whether these possibilities are plausible for the two data sequences in Figure 2.7, we see that the sequence in Panel B is more consistent with the first explanation (positive auto-correlation and an intervention effect) while the sequence in Panel A is more consistent with the second explanation (positive auto-correlation and no intervention effect). The JZS+AR Bayes factor takes these specific data patterns into account. Because the sequence in Panel A can be more easily explained by the null hypothesis than can the sequence in Panel B, the JZS+AR Bayes factor supports the alternative hypothesis of a true intervention effect more for the sequence in Panel B.

For the example data of Figure 2.1 the JZS+AR Bayes factor equals .70. Like the B_{jzs}, the B_{ar} favors the alternative hypothesis that there is a true intervention effect. However, B_{ar} is much closer to 1 than the B_{jzs} and thus contains much less evidence for the alternative. While the B_{jzs} favors the alternative with a factor of 3.7, the B_{ar} favors the alternative by only a factor of $1/.70 = 1.4$. This is not surprising given the data pattern in Figure 2.1. Although the means in the two phases differ, this decrease in scores can be easily explained by the negative trend (positive auto-correlation) in the data. This information is ignored by the JZS Bayes factor, which only takes the means and spread into account. However, the JZS+AR Bayes factor penalizes the evidence for the auto-correlation pattern in the data which results in a Bayes factor much closer to 1 in this case.

2.3.3 Bayes factors for trend and intercept differences

When there is only a stable difference in the level of the scores between two intervention phases, the mean difference appropriately summarizes the intervention effect. In this situation, B_{ar} is an appropriate measure of evidence for the mean change. However, often it is reasonable to assume that a sudden stable shift in scores is not the only pattern in the data. For example, there may be a gradual increase or decrease over time – that is, there may be a general trend in the data. Also, the trend after the intervention may be different from the trend before the intervention. In some cases the question of interest may be about changes in trends rather than in changes in means. In order to account for these possibilities, we extend the JZS+AR model to a model that also includes a general trend and a trend difference. Using this model we can answer questions about differences in both intercepts and trends between the two phases. We call this model the trend+AR (TAR) model. The primary difference between the TAR model and the JZS+AR model is the addition of two parameters: one for the general trend and one for the standardized difference between trends in the two phases. These

two parameters are analogous to the grand mean and mean level change in the JZS+AR model. With the new parameters, the model for each observation y_i is

$$y_i = \mu_0 + \sigma_z \delta x_i + \beta_0 t_i + \sigma_z \beta_1 x_i t_i + \epsilon_i,$$

where x_i is a dummy code as previously, and t_i is the time index, centered with the intervention at 0 (thus, $t_{n_1} = -0.5$, and $t_{n_1+1} = 0.5$). In this model, μ_0 and δ are the overall mean and standardized difference between intercepts, as in the JZS+AR model. Analogously, β_0 and β_1 are the general trend and standardized difference between trends. ϵ_i is the random error, which is modeled as in the JZS+AR model. Note that if $\beta_0 = \beta_1 = 0$ the TAR model reduces to the JZS+AR model.

Figure 2.8 shows the effects of changes in the parameters of the TAR model. Trend lines are fictitious true trend lines based on two "true" (errorless) observations in Phase 1 and two observations in Phase 2. For this visual demonstration, we assume $\sigma_z = 1$, so that the the unstandardized effects are the same as the standardized effects. In Panel A, the overall mean is $\mu_0 = 40$. This is the intersection point of the trend line with the dashed vertical line at centered time = 0; that is, the intercept. The general trend $\beta_0 = -15$, meaning that with one unit increase in time, the dependent variable y decreases by 15 units. Because there is no difference in the intercepts and trends between the two phases, δ and β_1 are zero. In Panel B, the trends in both phases are still equal and hence β_1 is still zero. However, the intercept is 20 units of y lower in the second phase than in the first phase, resulting in $\delta = -20$. In Panel C, the trends differ between the two phases but the intercepts do not. The trends in the first and second phases are -20 and -10 respectively, resulting in a trend difference β_1 of $-10-(-20) = 10$. In Panel D, the true trend lines show differences in both the intercepts and the trends, so that $\delta = -20$ and $\beta_1 = 10$.

The prior on nuisance parameters μ_0, β_0, σ_z^2 is analogous to the JZS+AR model:

$$p(\mu_0, \beta_0, \sigma_z^2) \propto \frac{1}{\sigma_z^2}.$$

As previously, we place Cauchy(r) priors on the parameters of interest, δ and β_1. And as before, the scaling factor r allows the adjustment of the prior distributions for different areas of study. Note that the interpretation of the δ and β_1 parameters in the TAR model differs from the interpretation of the δ parameter in the JZS+AR model, which requires new consideration of the r scales. Especially the standardized trend difference β_1 deserves some extra thought. Remember that a trend represents the amount of change in the dependent variable y with one unit increase in time (in our case, 1 time unit is the time between two successive observation points), and the β_1 parameter is the standardized difference between two of such trends. A complicating factor is that the amount of change per time unit depends on the number of observation points in a certain time interval. The smaller the amount of time between observation points, the smaller the change

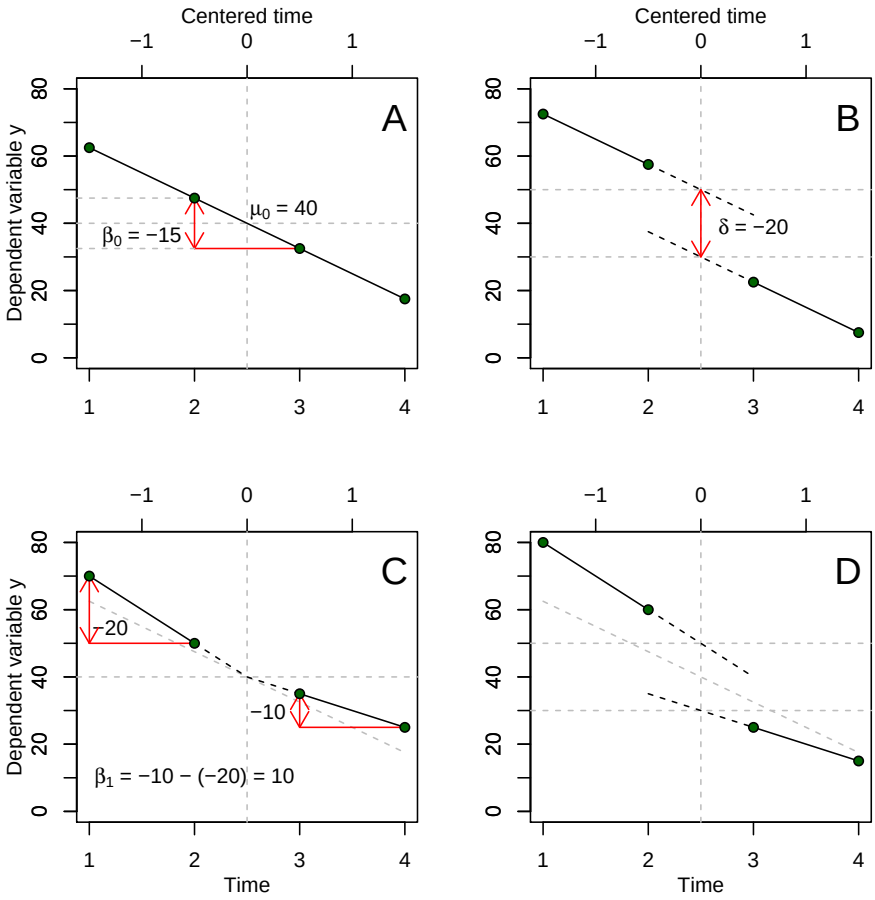

Figure 2.8: Visualization of TAR model parameters. A: $\mu_0 = 40$, $\delta = 0$, $\beta_0 = -15$, $\beta_1 = 0$; B: $\mu_0 = 40$, $\delta = -20$, $\beta_0 = -15$, $\beta_1 = 0$; C: $\mu_0 = 40$, $\delta = 0$, $\beta_0 = -15$, $\beta_1 = 10$; D: $\mu_0 = 40$, $\delta = -20$, $\beta_0 = -15$, $\beta_1 = 10$.

in y will be with one unit time increase. For instance, when the number of measurement points would be doubled within the same amount of time, keeping everything else the same, the expected amount of change per time unit would be half as large. Accordingly, the standardized difference β_1 between the trends would become half as large. When thinking about a reasonable r scale for the prior on β_1 one should thus take the time scale in the data into account.

Bayes factors can be computed for the intercept difference δ and the trend difference β_1, which together represent the intervention effect. There are a number of hypotheses worth comparing: first, there is the full null hypothesis, which we denote H_{00}:

$$\begin{aligned} \delta &= 0, \text{ and} \\ \beta_1 &= 0, \end{aligned}$$

and the full alternative hypothesis, which we denote H_{11}:

$$\begin{aligned} \delta &\sim \text{Cauchy, and} \\ \beta_1 &\sim \text{Cauchy.} \end{aligned}$$

We denote the Bayes factor of H_{00} against H_{11} as B_{i+t}, to indicate that it is a test of whether the intercept and trend differences are jointly null.

We can also specify hypotheses where either the intercept or trend differences, but not both, are null. For instance, the hypothesis H_{01} specifies that the intercept difference is null, while allowing the trend difference to be nonzero:

$$\begin{aligned} \delta &= 0, \text{ and} \\ \beta_1 &\sim \text{Cauchy.} \end{aligned}$$

Likewise, H_{10} specifies that the trend difference is null, while the intercept difference is nonzero:

$$\begin{aligned} \delta &\sim \text{Cauchy, and} \\ \beta_1 &= 0. \end{aligned}$$

We abbreviate the Bayes factor for H_{01} against H_{11}, which is a test of the intercept difference disregarding the trend difference, as B_{int}. Correspondingly, the test of H_{10} against H_{11}, a test of the trend difference disregarding the intercept difference, is denoted B_{trend}. The same techniques we used to compute the JZS+AR Bayes factors were used to compute the TAR Bayes factors (see Section 2 of the Supplement), and we provide easy-to-use functions in the BayesSingleSub R package available from the aforementioned website.

In order to show how the TAR Bayes factors extracted evidence from data, we simulated data with several levels of intervention effects, sample size, and autocorrelation, and computed B_{int}, B_{trend}, and B_{i+t}. The details of the simulations are described in Appendix B. We show only the Bayes factor for $\rho \sim \text{Beta}(1, 5)$ here; the effect of the prior for ρ on the TAR Bayes factors is similar to its effect

on the JZS+AR Bayes factor. We set the r scales for the Cauchy priors to 1, but as discussed above, researchers can choose different r scales based on what is reasonable for the time scales and effects in their data. Different r values will not change the patterns in the simulations, only their magnitude.

Figure 2.9 shows B_{trend} and B_{int} as a function of the least squares estimates of β_1 and δ (ignoring auto-correlation), respectively, for different sample sizes per phase N. As before, the solid lines are nonparametric regression lines and the dots show the variation in the Bayes factors due to the patterns in the data. The bottom row of Figure 2.9 clearly shows a similar pattern for B_{int} as for B_{ar}. When sample size is small, the B_{int} remains close to 1, indicating that the data contain little information about the difference between intercepts. When sample size gets larger, the information in the data increases and the B_{int} deviates more from 1.

The top row of Figure 2.9 shows that, as would be expected, larger data sets provide more information about trend differences than smaller data sets. In addition, it looks like the B_{trend} is much more responsive to changes in observed trend differences than the B_{int} and B_{ar} are to changes in observed intercept differences. However, it is hard to compare the B_{trend} with the B_{int} and B_{ar} in this way, since it is not clear how an observed trend difference of a certain size relates to an observed intercept difference of the same size. That is, these numbers may have different meanings in terms of effect size.

The joint Bayes factor B_{i+t} behaves as one would expect based on Figure 2.9, as a function of β_1 and δ: B_{i+t} favors the alternative more as the least-squares estimates of β_1 and δ become more extreme. Interestingly, the information in the two Bayes factors B_{trend} and B_{int} is nearly, though not entirely, independent. Figure 2.10 shows the relationship between the joint Bayes factor B_{i+t} and the product of B_{int} and B_{trend}. The points lie near the diagonal, indicating near-independence[7]. We suspect this is related to the fact that the sums of squares for the two effects would be independent in the absence of serial-autocorrelation, and are nearly so when autocorrelation is moderate. In addition, B_{i+t} appears to have a slight bias toward the null hypothesis, relative to the product of B_{int} and B_{trend}. This bias is expected and reflects the Bayes factor's natural penalty for the greater flexibility of the general model with both intercept and trend differences.

For the example data in Figure 2.1, B_{trend} equals 4.0, B_{int} equals 1.6, and B_{i+t} equals 5.3. That is, in contrast to the JZS+AR Bayes factor which slightly favored the alternative hypothesis of a mean difference with a factor of 1.4, all Bayes factors from the TAR model favor the null hypothesis of no intervention effect. These contrasting findings are not surprising when we consider the specific hypotheses that are compared by the different Bayes factors. The B_{ar} considers the difference in means while taking the auto-correlation into account. As the B_{ar} made clear, the data show evidence for a mean difference between the two

[7] There is undoubtedly also estimation error contributing to the variance in the Bayes factors, but Morey et al. (2011) showed that with the estimation technique used here, estimation error is likely to be small.

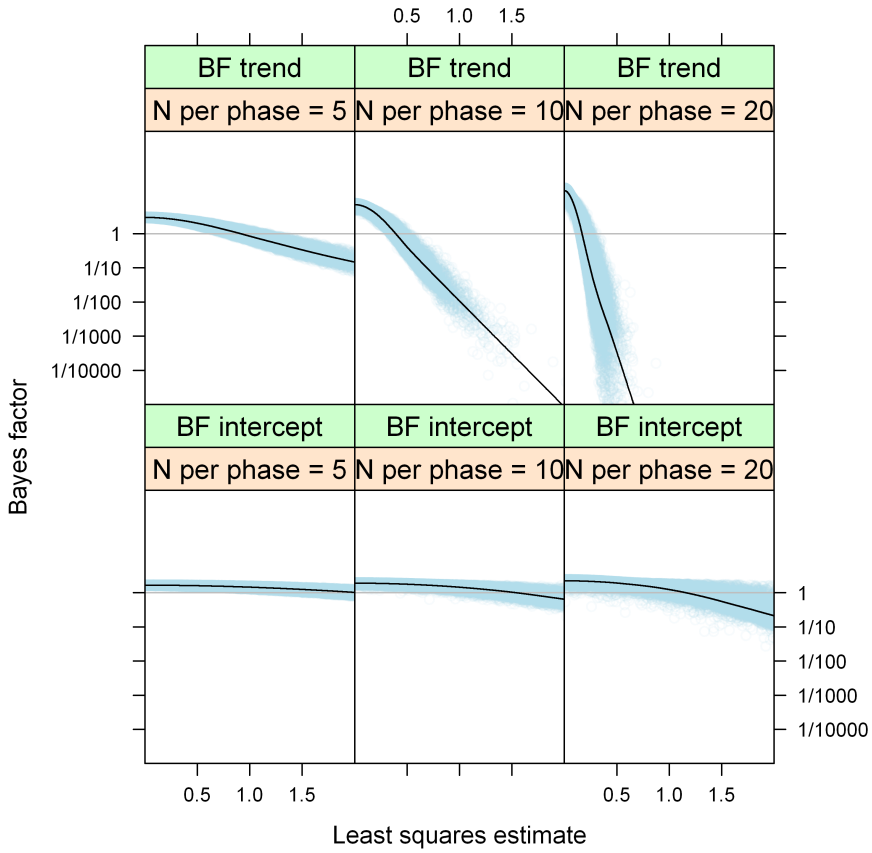

Figure 2.9: B_{trend} (top row) and B_{int} (bottom row) as a function of least squares estimates of β_1 and δ, respectively, for several sample sizes; solid lines are non-parametric regression lines.

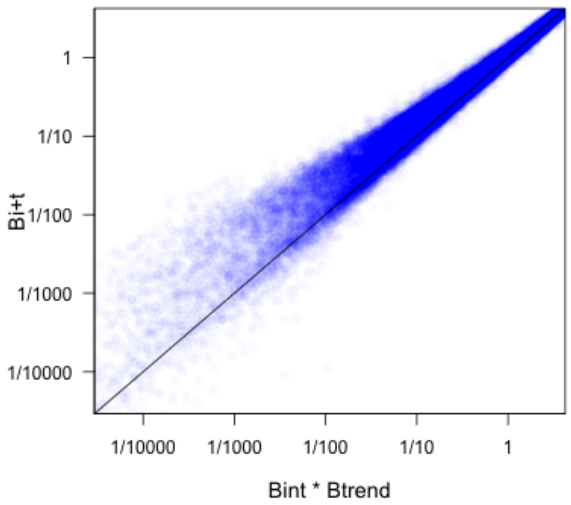

Figure 2.10: B_{i+t} versus $B_{trend}B_{int}$. The diagonal line represents independence of the two Bayes factors.

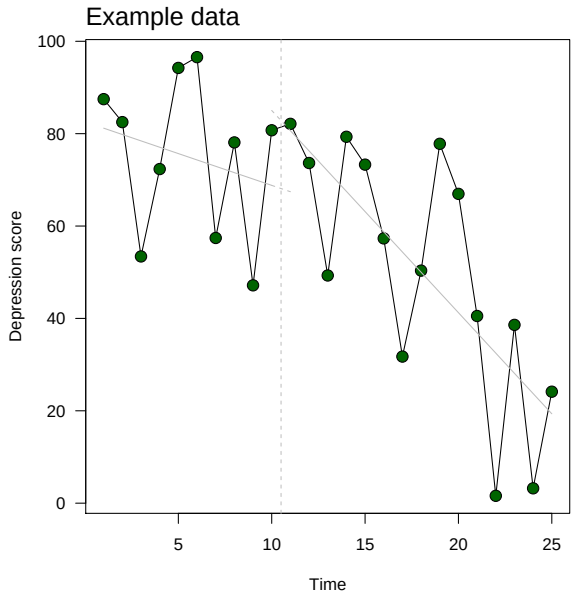

Figure 2.11: Example data with least squares estimates of trend lines.

phases suggesting, according to this model, an intervention effect. However, the JZS+AR model does not consider the possibility that this mean difference is due to a downwards overall trend in the data. The TAR model, on the other hand, admits the possibility of a general downward trend and focusses on differences in trend lines. Although the data do show some differences in intercept and trend, as shown in Figure 2.11, there is not enough evidence that these differences reflect true intervention effects, rather than random fluctuation around a generally decreasing trend.

2.3.4 Estimation of effect sizes and credible intervals

Although hypothesis testing using Bayes factors is a useful way to weigh the relative evidence for the null and alternative hypotheses, hypothesis testing is not the only way to explore the evidence for an intervention effect. The use of interval estimates, such as credible intervals and confidence intervals, has often been advocated as a supplement to hypothesis testing (Reichardt and Gollob, 1997; Rouder and Morey, 2005), and sometimes as a replacement (Loftus, 1996; Schmidt and Hunter, 1997). In this section, we show how point and interval estimates can be naturally obtained using the TAR model.

In our description of the general Bayes factor technique, we described how prior distributions are used as weighting functions to compute the average, or marginal, likelihood for a model. To compute a marginal likelihood, it is necessary to average over the uncertainty in all unknown parameters. However, if we were interested in examining the likely values, given the data, for a particular parameter — say, the trend difference β_1 — we might average over all parameters *except* the parameter of interest under the alternative hypothesis. What is left is a probability distribution representing the uncertainty in the parameter of interest after taking the data into account, assuming that an effect exists. This probability distribution is called a *marginal posterior distribution*; *marginal*, because uncertainty in all other parameters has been averaged out, and *posterior* because it represents uncertainty after observing the data (as opposed to the prior, which represents uncertainty before observing the data). When the sample size is high, and thus the data contain much information, the variance of the posterior will be low and the posterior distribution will be largely determined by the data. When the sample size is low, the data contain less information, and the posterior distribution is more determined by the prior distribution.

The same techniques we used to compute the trend and intercept Bayes factors can be used to estimate the posterior distributions for intercept and trend differences. For ease of interpretability and comparison with the plots, we show posterior distributions on the unstandardized intercept and trend differences $\alpha_1 = \sigma_z\delta$ and $\alpha_2 = \sigma_z\beta_1$, but posterior distributions can be estimated for standardized effects or for any other model parameter. Examination of the posterior distributions for α_1 and α_2 shows what values of these parameters are plausible in light of the data, and which are not. Figure 2.12 shows the marginal posterior distributions of α_1 and α_2 for the example data in Figure 2.1. The means of

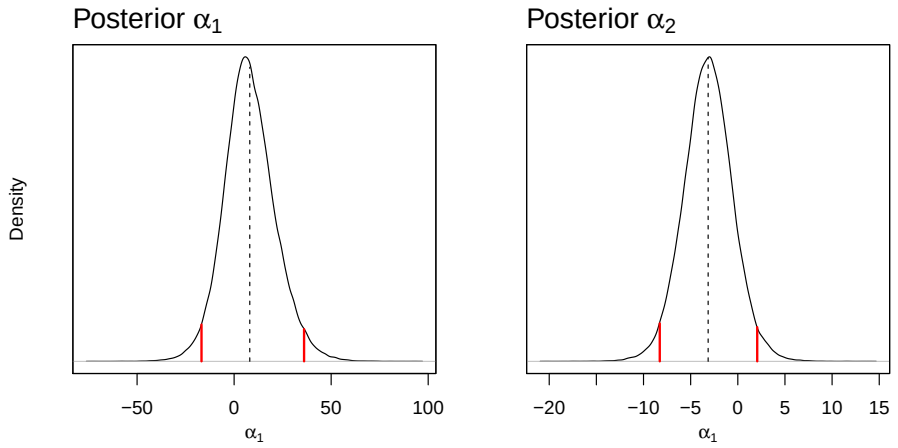

Figure 2.12: Posterior distributions for the intercept difference $\alpha_1 = \sigma_z \delta$ and the trend difference $\alpha_2 = \sigma_z \beta_1$; dashed vertical lines in the middle represent means, vertical lines in the tails of the distributions represent bounds for 95% credible intervals.

the posterior distributions, shown as dashed vertical lines in the middle of the posteriors, can be used as point estimates of the true effect size. The posterior means of α_1 and α_2 are 8.1 and -3.1, respectively. That is, based on the data and the prior expectations about the parameters, the intercept is estimated to be eight points larger and the trend is estimated to be three points steeper in the second phase than in the first phase. This is in line with the data patterns shown in Figure 2.11.

We might also be interested in calculating interval estimates, such as confidence intervals. In Bayesian statistics, the most common form of interval estimate is the 95% *credible interval*, which is an interval containing 95% of the posterior density, typically chosen to exclude 2.5% of the posterior area in each tail. A credible interval can thus be interpreted as an interval in which there is a 95% chance that the true parameter lies, when the prior distributions and data are taken into account. In Figure 2.12, the bounds for the 95% credible intervals for α_1 and α_2 are indicated with vertical line segments in the tails of the posterior distributions. The figure shows that with 95% probability, α_1 lies between -17 and 36 and α_2 lies between -8.3 and 2.1. Both credible intervals have a wide spread, indicating that there is a large amount of uncertainty about the amount of intercept and trend change. Also, both intervals indicate that the true change could be either positive or negative; at sample sizes this low in the presence of autocorrelation, it is difficult to assess even the sign of the effect, if it exists.

2.3.5 Extension of Bayes factor to clinical significance

Examining posterior distributions also allows the determination of whether a given effect is "clinically significant" (Jacobson and Truax, 1991; Cohen, 1994; Wellek, 2003). Using Bayes factors, it is possible to obtain evidence that the intervention effect is nonzero, but credible intervals may indicate that the effect is likely to be very small. In this case, the fact that the evidence for an effect is strong is not particularly interesting. Clinicians are often interested in interventions that will be more than negligibly effective, and thus may find that testing for instance the point null hypotheses that $\delta = 0$ is not useful. Instead, it may be more desirable to choose a positive cut-off c such that standardized effect sizes greater than c (or less than $-c$) are considered clinically significant, and effect sizes smaller than c in magnitude are not. A Bayes factor may then be computed to test the null hypothesis that $|\delta| < c$ against the alternative hypothesis that $|\delta| > c$ (and similarly for β_1). These hypotheses are visualized in Figure 2.13. Morey and Rouder (2011) describe methods for extending the JZS Bayes factor to cases in which the null hypothesis is a range of small effect sizes rather than that $\delta = 0$. We applied these methods to extend our Bayes factors to Bayes factors for interval null hypotheses.

We will illustrate these Bayes factors with the example data and the TAR model. Based on experience, theory, and other information, we may consider $|\delta|$ values up to .2 and $|\beta_1|$ values up to .1 practically irrelevant. Hence our null hypotheses of interest would be that the standardized difference in intercepts is between $-.2$ and .2 and that the standardized difference in trends is between $-.1$ and .1. The alternative hypotheses would be that the difference in intercepts and trends, respectively, exceed these bounds. The extended Bayes factor for the intercept difference compares the interval null hypothesis that $|\delta| < .2$ to the interval alternative hypothesis that $|\delta| > .2$. Similarly, the extended Bayes factor for the trend difference tests $|\beta_1| < .1$ against $|\beta_1| > .1$. For the example data, the interval Bayes factors for the intercept and trend differences equal and 1.8 and 5.4, respectively. These Bayes factors are similar to the corresponding point null Bayes factors, which were 1.6 and 4.0, respectively. This similarity does not always hold in practice, however. Especially when the observed effect size falls within the bounds of the interval null hypothesis and the sample size is large, the interval and point null Bayes factors may diverge.

2.4 Discussion

In the preceding sections, we have developed several Bayes factor statistics that are useful for evaluating evidence for competing hypotheses in single-subject research. The Bayes factors developed here are appropriate for interval- and ratio-scaled data from participants measured repeatedly over time, at roughly equal time intervals. The JZS+AR Bayes factor allows for testing an intervention effect where no trend exists, and the three TAR Bayes factors allow for testing intervention effects on intercepts and trends. In general, the Bayes factor, being the

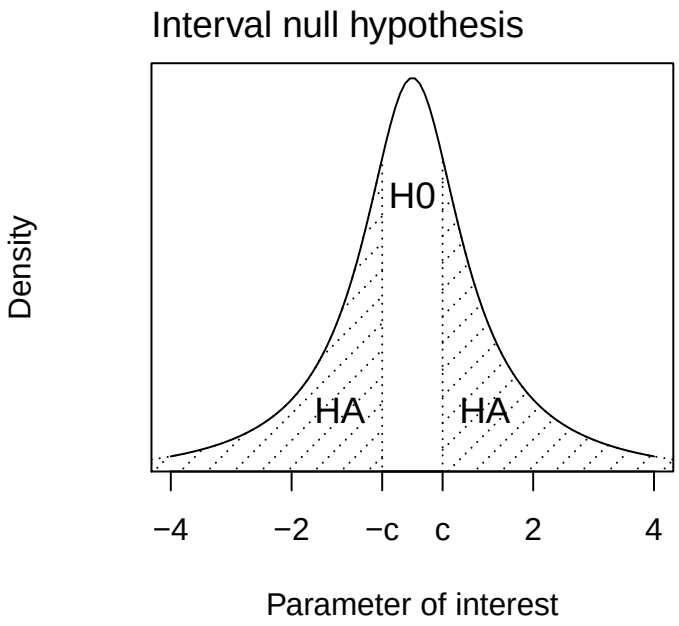

Figure 2.13: Interval null and alternative hypothesis.

degree to which a rational person observing the data should adjust their beliefs (expressed as odds) in favor of one or the other hypothesis, represents a principled measure of evidence.

Like all statistical models, the models developed here rely on assumptions, such as homogeneity of variance and normality. Ideally these assumptions should be checked in order to see whether the model is reasonable for the data, and there are several ways to check these assumptions in large data sets. However, the small data sets usually encountered in single-subject research provide little information about the distribution of the residuals, which makes checking assumptions difficult. But even if the assumptions are violated, the results of an inferential technique can still be informative, with appropriate caveats. Still, when applying a parametric model like the JZS+AR or TAR model, it should be kept in mind that the usefulness of the model is conditional on the assumptions made by the model. This is not specific to the models presented in this paper, of course, but holds for any inferential technique.

Bayes factors represent a different type of data analysis tool from what most single-subject researchers are probably familiar with. We have thus far avoided discussing alternative techniques, in order not to distract from our main goal: introducing our Bayes factor statistics. The ITSA model (McDowall et al., 1980), for instance, bears some similarity with our JZS+AR model in that it models both an ARIMA process and an intervention effect. There are differences, however, in the way parameters are estimated between ITSA and the JZS+AR model. These differences arise from the differences in perspective between classical statistics and Bayesian statistics. Below, we discuss in more detail some of the fundamental differences between Bayesian techniques and more traditional statistical techniques.

2.4.1 Bayesian methods versus null hypothesis significance testing

In this paper, we have taken the Bayesian point of view for granted. Bayesian statistics, although it has become standard among statisticians, has not had as much success among psychologists. There are several reasons for this. Perhaps the most important one is training: Bayesian techniques are not yet taught in standard data analysis classes. More commonly taught are so-called "classical" techniques, such as null hypothesis significance testing (NHST). NHST techniques, such as Student's t test and ANOVA for group designs and interrupted time series analysis and permutation tests for single-subject designs, dominate the psychological literature. Although classical techniques have been well-represented in the single-subject literature, Bayesian techniques have not. One of our goals with this paper is to rectify this situation. Some researchers who have been trained only in the use of classical methods are hesitant to use alternative Bayesian techniques. We believe that this hesitation is unwarranted.

The major reason we advocate Bayesian techniques is that they offer something that NHST cannot: the possibility to measure evidence (Good, 1985). Tra-

ditional NHST techniques are based on the idea of controlling error rates. In null hypothesis significance tests, a null hypothesis is posited, and a test statistic is computed which quantifies the degree to which the data are inconsistent with the null hypothesis. The probability of obtaining a more extreme test statistic under the null hypothesis, called a p value, is computed. If this p value is low enough, typically less than .05, then the null hypothesis is rejected in favor of the alternative hypothesis. Because the p value is computed assuming the null hypothesis is true, this ensures that if indeed the null hypothesis is true, we would only incorrectly reject it 5% of the time. This type of error is called a Type I error.

Although the p value is often described as a measure of evidence against the null hypothesis, it is not. The same p value — say, .05 — may represent strong evidence against the null hypothesis, such as when the sample size is low or, seemingly paradoxically, strong evidence against the alternative hypothesis when the sample size is very large (Lindley, 1957; Sellke et al., 2001). As the sample size increases, the p values will tend to grow smaller. For very large sample sizes, a marginal p value like .05 would be unexpected under any reasonable alternative, but would not be as uncommon under the null hypothesis. Thus, although a p value of .05 would traditionally be seen as "sufficient" evidence to reject the null hypothesis, under some conditions it is evidence *for* the null. This seemingly paradoxical behavior of p values is known in the statistical literature as the Lindley paradox (Lindley, 1957). The apparent paradox arises because NHST uses the null hypothesis as a default, and does not compare the fit of the null hypothesis to the fit of any reasonable alternatives.

Because a given p value corresponds to different levels of evidence depending on the sample size, it cannot be used as a measure of evidence (Berger and Sellke, 1987). This is not in itself a problem; NHST can be used to construct tests with a known Type I error rate. If control of Type I error rate is desired, then NHST provides a method for doing so. However, in science, it is often necessary to evaluate evidence. In scientific practice, p values are often used a proxy for evidence, although no clear rationale exists for doing so, and good reasons exist for not doing so (see Wagenmakers et al., 2008, for a review). In contrast, Bayes factors are the degree to which relative belief in two hypotheses, quantified as odds, should change in light of the data. This corresponds to a very straightforward definition of evidence, and thus, we would argue that Bayes factors are ideal for scientific communication. This is not to say that there is no place for other statistical techniques such as NHST; however, we believe that Bayes factor should play a dominant role when measures of evidence are desired.

It might be objected that a Bayes factor only provides a measure of evidence in light of the priors chosen. Indeed, as we have shown above, Bayes factors depend on the prior distributions. This is not, however, surprising, and we do not view it as a weakness. Scientific evidence must always be, and *should* always be, interpreted in a context. Scientific researchers regularly evaluate evidence when they read studies in the scientific literature; this is always done in the context of what the researcher knows about that literature. Likewise, the priors used to

compute a Bayes factor provide a context (Berger and Berry, 1988). The priors we selected for the standardized intervention effects and the lag 1 autocorrelation ρ are informed by the scientific literature (Beeson and Robey, 2006; Fox, 2008, chap. 16; Jeffreys, 1961; Matyas and Greenwood, 1997; Parker et al., 2005, 2007; Parker and Vannest, 2009; Zellner and Siow, 1980) and experience in data analysis. Measures of evidence computed using reasonable priors will be reasonable, and will be more likely to garner agreement among researchers; measures of evidence computed using unreasonable priors will be unreasonable, and unlikely to be taken seriously. In some situations, for instance in new research areas, it may be harder to come up with one reasonable prior distribution and several prior distributions may be proposed. Substantive conclusions can then be reached under each of the several priors specified. When conclusions are similar across different, but reasonable, prior distributions, the substantive conclusions are credible in spite of differences across assumptions. Researchers using the Bayes factor methods should report the prior settings used, allowing other researchers to evaluate whether the priors are reasonable.

The debate between advocates of NHST and advocates of Bayesian techniques in psychological methods is an ongoing one. Bayesian analysis has other advantages beyond the ability to compute evidence, such as ease of interpretation, ability to accumulate evidence for null and alternative hypotheses, a solid axiomatic foundation, and many others. There are several excellent resources for readers interested in further arguments for Bayesian statistics, including Edwards et al. (1963) and Jaynes (2003). Jaynes (1986) provides a history of Bayesian logic, and Wagenmakers et al. (2008) provide a recent, comprehensive review of the arguments for the use of Bayesian techniques. It is our hope that readers will find that Bayesian statistics provides useful tools for learning from their data.

2.4.2 Required number of data points

Many authors have shown estimation problems of the auto-correlation ρ when sample size is small (Solanas et al., 2010), which causes problems in classical inferential analyses. Positive auto-correlation reduces the information contained in the data, and the larger the auto-correlation, the more the information in the data is reduced. Thus, when the estimate of the auto-correlation is inaccurate, the information in the data is under- or overestimated resulting in inappropriate inferences.

Fortunately, in the Bayesian approach a small data set is less problematic in the presence of auto-correlation. This is because the prior distribution on ρ is spread out over a range of plausible values for ρ and in this way takes into account the uncertainty in ρ. This contrasts with the single point estimate of ρ used in the classical approach. When the number of data points is small, the data contain little information about the true level of auto-correlation and the posterior is largely determined by the prior distribution for ρ. This is shown in Figure 2.14A and 2.14C, which show a short data sequence ($N = 10$) with corresponding prior (dashed line) and posterior distribution (solid line) for ρ, from the TAR model.

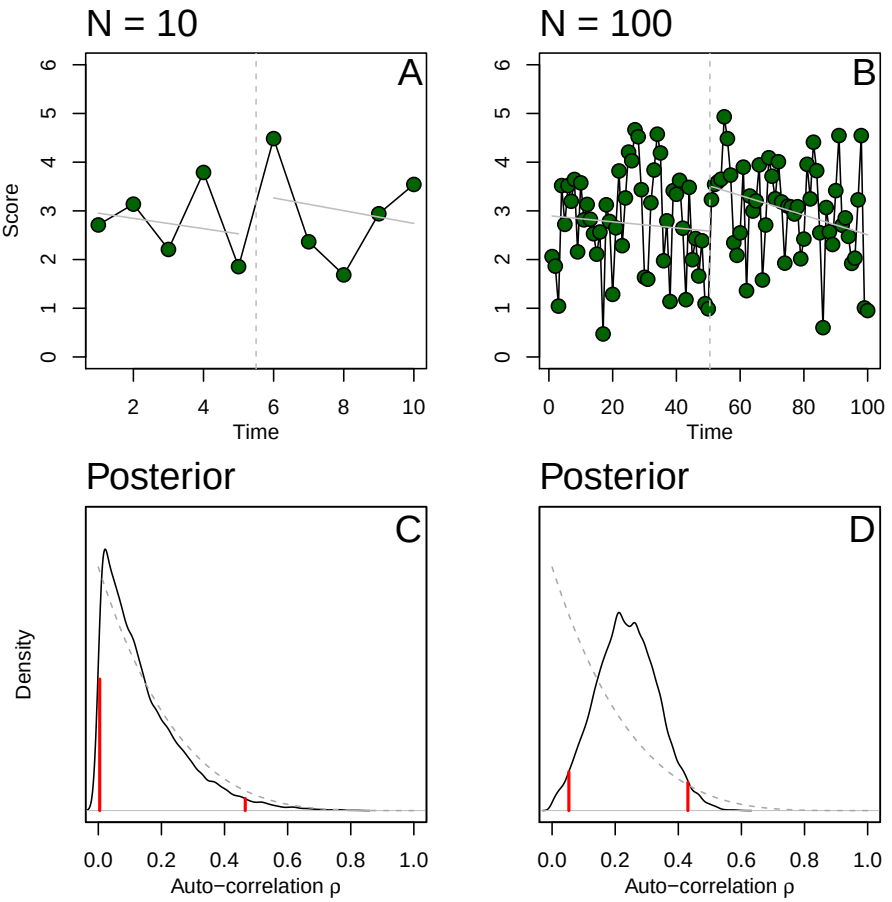

Figure 2.14: First row: small data sequence (total $N = 10$) and large data sequence (total $N = 100$) with a true auto-correlation of .3. Second row: corresponding posterior distributions (solid lines) and prior distributions (dashed lines) for the auto-correlation ρ; vertical lines in the tails of the distributions represent bounds for 95% credible intervals.

The prior and posterior distributions are almost similar for this small data set. Note however that the prior distribution represents a weighting distribution of *a priori* likely values for the auto-correlation. A distribution close to the prior distribution is thus the reasonable posterior distribution when the data contain almost no information, providing a better representation of uncertainty about the parameter ρ than a single point estimate based on the data.

When the number of data points is larger, the posterior for ρ is more determined by the data. This is shown in Figure 2.14B and 2.14D, which show a large data sequence ($N = 100$) with corresponding prior and posterior distribution for ρ. For this larger data set the posterior for ρ is pulled more towards the true value of .3 and the distribution is narrower because there is less uncertainty about ρ. This increases the evidence contained in the Bayes factors. Although more evidence is obviously desirable in inference, it does not mean that the analysis based on the small data set is wrong as long as the prior distributions are reasonable. This contrasts with the classical approach, where a small data set would result in a single, volatile estimate of the auto-correlation, which would distort inferences.

One last sample-size consideration is power. Readers familiar with classical methods may wonder about the power of the procedure outlined here. Power considerations are relevant in classical methods, which condition on a true model and focus on Type I and Type II error rates. However, in Bayesian statistics, the conditioning is reversed - statistics are conditioned on the data, and the evidence for particular hypotheses *given the data* is computed. While the notion of power is meaningful in the context of classical methods, it is not as meaningful with Bayesian methods. Researchers designing experiments may wonder how sample sizes and evidence are related; our simulations should provide a rough guideline for how much evidence researchers can expect their data to contain, given particular sample sizes and summary statistics.

2.4.3 Conclusions

In this paper, we have developed formal statistical tests for testing for intervention effects in single-subject data. The JZS+AR Bayes factor method we describe allows researchers to evaluate the evidence for the null hypothesis of no mean difference relative to the evidence for a mean difference between phases. The TAR Bayes factors evaluate the evidence for trend and intercept differences. We recommend that single-subject researchers use formal statistical methods, such as our TAR Bayes factors, alongside visual inspection of their data. The JZS+AR and TAR Bayes factors provide measures of evidence for intervention effects that take into account both random variation and serial dependencies. R functions to compute the Bayes factors described in this paper are included in the BayesSingleSub R package, which can be installed from within R. An R guide to use these functions is available in the online Supplement, Section 1.

2.5 Appendix A

The Bayes factor is not the same as the posterior odds of the null over the alternative hypothesis. The Bayes factor is the extent to which a rational person will modify their prior beliefs, expressed as relative odds, in light of the data.

One way to express the Bayes factor is

$$B = \frac{p(\text{data} \mid H_0)}{p(\text{data} \mid H_1)}.$$

This way of expressing the Bayes factor makes explicit the fact that the Bayes factor is the ratio of the likelihoods of the data under each hypothesis, after the uncertainty in all the parameters has been averaged out.

Using Bayes theorem yields a way to relate this to prior and posterior odds for the two hypotheses. By Bayes theorem,

$$p(\text{data} \mid H_0) = \frac{p(H_0 \mid \text{data})p(\text{data})}{p(H_0)}$$

and likewise for H_1. By dividing both sides of the expression in this equation by the corresponding equation for H_1, we obtain

$$\frac{p(\text{data} \mid H_0)}{p(\text{data} \mid H_1)} = \frac{p(H_0 \mid \text{data})}{p(H_1 \mid \text{data})} \Big/ \frac{p(H_0)}{p(H_1)}$$

The term in the numerator on the right-hand side is called the *posterior odds*, and represents relative belief in two hypotheses after observing the data. The term in the denominator on the right-hand side is called the *prior odds*, and represents the relative belief in two hypotheses before observing the data. This way of expressing the Bayes factor shows that the Bayes factor corresponds to the change in odds from prior to posterior based on the data by a rational observer. Individuals holding different ideas about the prior odds can use the Bayes factor to adapt their idiosyncratic prior odds to their idiosyncratic posterior odds. If we give equal weight to the null and alternative a priori, the prior odds are 1 and the posterior odds are numerically equal to the Bayes factor. We stress, however, that positing particular prior odds is not necessary for interpreting the Bayes factor; the Bayes factor is insensitive to the prior odds, and does not require them for interpretation.

2.6 Appendix B

2.6.1 Simulations for JZS+AR model

Data were generated in the R statistical environment (R Development Core Team, 2009). For the JZS+AR model data were generated according to

$$y_i = \mu_0 + \sigma_z \delta x_i + \epsilon_i,$$

where y_i represents the data value at time point i, x_i indicates phase change coded -.5 before and .5 after the intervention, and ϵ_i represents random error at time point i. In this way μ_0 is the overall mean and δ is the standardized mean difference between the data in the baseline and post-intervention phases. The random errors ϵ_i were generated according to an AR(1) process:

$$\epsilon_i = \rho \epsilon_{i-1} + z_i,$$

with ρ the lag 1 auto-correlation and z_i the random shock at time point i. For each data set the number of random errors generated equaled 50 plus the sample size N, whereafter the first 50 errors were eliminated. This so called burn in sequence of 50 errors ensured that the last N errors of the sequence were not affected by the first (arbitrary) value of the sequence.

The μ_0 parameter was set to zero, which does not affect the outcomes. The δ parameter was set at 0, .5, and 2. However, by random sampling this could result in observed Cohen's d values as large as 10, due to the low sample sizes used in this simulation. Sample sizes per phase were set at 5, 10, and 20. About 10 data points per phase is the maximum number possible in practice (Parker and Brossart, 2003; Parker and Hagan-Burke, 2007; Parker and Vannest, 2009). We included a sample size of 20 per phase to show what happens when the data set is doubled. The true auto-correlation ρ ranged from 0 to .9, in steps of .3. The standard deviation of the random shocks z_i was set at 1. Together this resulted in a 3 (Cohen's δ) $\times$ 3 (sample size) $\times$ 4 (auto-correlation) simulation design with 36 cells. We performed 500 repetitions per cell producing a total of $36 \times 500 = 18,000$ different simulated data sets.

For each simulated data set, parameters of the JZS+AR model were estimated using Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990), which is explained in Section 2 of the Supplement. The AR Bayes factor was estimated with the Savage-Dickey density ratio, which relates the Bayes factor to the ratio of the marginal prior distribution to the marginal posterior distribution at the restriction $\delta = 0$ within the unrestricted model, see Dickey and Lientz (1970); Morey et al. (2011), and Section 2 of the Supplement. The chains of the Gibbs sampler consisted of 5000 iterations. Because of the fast convergence and the large number of iterations, no iterations where discarded. Convergence was checked by visual inspection of the chains and by comparing the Bayes factors resulting from the Savage-Dickey density ratio with the Bayes factors resulting from the Monte Carlo estimate for several data sets. Convergence was observed in

all cases. For comparison, three different priors on the auto-correlation parameter were used: beta$(1, 1)$, beta$(1, 5)$, and beta$(1, 15)$. See Section 2 of the online supplement for details.

2.6.2 Simulations for the TAR model

For the TAR model, data were generated according to

$$y_i = \mu_0 + \sigma_z \delta x_i + \beta_0 t_i + \sigma_z \beta_1 x_i t_i + \epsilon_i,$$

where y_i, x_i, t_i, and ϵ_i are defined as before (see Section 2.3.3). The random errors ϵ_i were generated according to the same AR(1) process as for the JZS+AR model.

As for the JZS+AR model, μ_0 was set to zero. The parameter δ was set at 0, .5, and 2, and β_0 and β_1 were both set at 0, .1, and .25. The sample sizes per phase were set at 5, 10, and 20, as before. The true auto-correlation ρ ranged from 0 to .6, in steps of .3. We did not include true auto-correlations of .9 in these simulations because it is an unlikely amount of auto-correlation for empirical data and the simulations for the JZS+AR model already demonstrated how the Bayes factor is affected by it. The standard deviation of the random shocks z_i in the AR(1) error model was again set at 1. Together this resulted in a 3 (δ) $\times$ 3 (β_0) $\times$ 3 (β_1) $\times$ 3 (sample size) $\times$ 3 (auto-correlation) design with 243 cells. We performed 500 repetitions per cell producing a total of $243 \times 500 = 121,500$ different simulated data sets.

Procedures for estimating parameters of the TAR model were similar to those for the JZS+AR model. However, because the effect of the different beta priors for ρ had already been investigated for the JZS+AR model, we only used the beta$(1, 5)$ prior in these simulations. Also, $10,000$ Gibbs sampler iterations were used instead of 5000, due to the short time in which the simulations could be run for the JZS+AR model.

2.7 Online Supplement

This document provides example R code demonstrating how to use the BayesSingleSub package (Section 1), and the technical details for the sampling routines (Section 2).

2.7.1 Tutorial for computing de Vries and Morey's Bayes factors

Here, we show how to compute the Bayes factors B_{ar}, B_{trend}, B_{int}, and B_{t+i}, and how to obtain and plot the posterior distributions of the model parameters.

First, download the R statistical environment from `http://cran.r-project.org/` and install the BayesSingleSub package using the R command:

```
> install.packages("BayesSingleSub")
```

Then, load the BayesSingleSub package with the `library()` function:

```
> library(BayesSingleSub)
```

For the purposes of this demonstration, we compute the Bayes factors for the data shown in Figure 1 of the manuscript. We first define the data and the number of observations in the pre- and post-treatment phases:

```
> data = c(87.5, 82.5, 53.4, 72.3, 94.2, 96.6, 57.4, 78.1,
           47.2, 80.7, 82.1, 73.7, 49.3, 79.3, 73.3, 57.3,
           31.7, 50.4, 77.8, 67, 40.5, 1.6, 38.6, 3.2,
           24.1)
> n1 = 10
> n2 = 15
```

For convenience, we divide the data before and after the intervention into separate vectors:

```
> ypre =  data[1:n1]
> ypost = data[n1 + 1:n2]
```

The logarithm of the JZS+AR Bayes factor B_{ar} can be obtained by using the `ttest.Gibbs.AR()` function, and the logarithm of the TAR Bayes factors B_{int}, B_{trend}, and B_{i+t} by using the `trendtest.Gibbs.AR()` function:

```
> logBAR = ttest.Gibbs.AR(ypre, ypost, iterations = 10000,
          return.chains = FALSE, r.scale = 1,
          betaTheta = 5, sdMet = 0.3)
> logBTRENDS = trendtest.Gibbs.AR(ypre, ypost,
            iterations = 10000, return.chains = FALSE,
            r.scaleInt = 1, r.scaleSlp = 1, betaTheta = 5,
            sdMet = 0.3)
```

which will compute the Bayes factors while setting the r scales of the Cauchy priors to 1, and the parameter b of the beta priors on ρ to 5. These are the default for the `ttest.Gibbs.AR()` and `trendtest.Gibbs.AR()` functions.

The first and second arguments of both functions are the series of observations in Phase 1 and Phase 2, respectively. The `iterations` argument controls the number of Gibbs sampler iterations; more iterations will increase the accuracy of the estimate of the Bayes factor. The accuracy of the estimate can be checked by comparing the estimate from the Gibbs sampler with the Monte Carlo estimate, discussed below. Substantial disagreement implies that the Gibbs sampler has not yet converged and the number of iterations should be increased. Setting the `return.chains` argument to `FALSE` ensures that the MCMC chains are not returned. They can be returned if they are needed, as we show below. The values for r can be changed by changing the `r.scale` argument of the `ttest.Gibbs.AR()` function and by changing the `r.scaleInt` (for the intercept differences) and `r.scaleSlp` (for the trend differences) arguments of the `trendtest.Gibbs.AR()` function.

In both functions the value for b can be changed by changing the `betaTheta` argument. If desired, r and b can be changed a few times and resulting Bayes factors can be compared. Finally, the "acceptance rate" reported by the function is an index of the quality of the MCMC sampling of ρ (the Metropolis-Hastings acceptance rate; Hastings, 1970; Ross, 2002). This number should be between .25 and .5 for most efficient estimation. If needed, the acceptance rate can be increased or decreased by decreasing or increasing, respectively, the `sdMet` argument, but the default setting should suffice for almost all analyses. For more information about a function's arguments, see the R help files for the corresponding function (e.g., `help("ttest.Gibbs.AR")`).

The `logBAR` variable now contains an estimate of the logarithm of the JZS+AR Bayes factor, and the `logBTRENDS` variable contains the logarithm of the three trend Bayes factors. We can exponentiate these log Bayes factors to obtain the Bayes factors:

```
> logBAR

[1] -0.3615208

> logBTRENDS

 logbf.i+t logbf.trend logbf.int
  1.719709    1.426854 0.4989952

> exp(logBAR)

[1] 0.6966161

> exp(logBTRENDS)

 logbf.i+t logbf.trend logbf.int
  5.582905    4.165572  1.647066
```

Every time the code above is run, the values will be slightly different, due to the random nature of MCMC estimation. However, with sufficient iterations (typically 2,000 or greater, in this application) the estimate should be consistent across calls to the `ttest.Gibbs.AR()` and `trendtest.Gibbs.AR()` functions.

If we wish to examine the posterior distribution for any parameter or the interval null Bayes factors for δ and β_1, we may do so by first calling `trendtest.Gibbs.AR()` function with the `return.chains` argument set to true and the bounds under the null hypotheses defined[8]:

```
> output.trend = trendtest.Gibbs.AR(ypre,ypost,
                iterations = 10000,
                return.chains = TRUE, r.scaleInt = 1,
                r.scaleSlp = 1, betaTheta = 5, sdMet = 0.3,
                intArea = c(-0.2, 0.2),
                slpArea = c(-0.1, 0.1))
```

The interval null Bayes factors are only returned if the `return.chains` argument is set to true. The default bounds under each of the null hypotheses changed by changing the `areaNull` argument of the `ttest.Gibbs.AR()` function and changing the `intArea` (for the intercept) and `slpArea` (for the trend) arguments of the `trendtest.Gibbs.AR()` function. Note that the chains contain *unstandardized* parameters, and the interval null Bayes factors are based on standardized effect sizes.

The variable `output.trend` now contains four components: `logbf`, which contains an estimate of the Bayes factor(s), `chains`, which contains the MCMC chains, `acc`, which contains the Metropolis-Hastings acceptance rate described above, and `logbfArea`, which contains the Bayes factor(s) for interval null hypotheses.

```
> logIntervalNullBF.trend = output.trend$logbfArea
> chains.trend = output.trend$chains
```

The variable `logIntervalNullBF.trend` contains the logarithm of the interval-null Bayes factors for the intercept and slope, respectively. Exponentiating the log Bayes factors gives the Bayes factors:

```
> logIntervalNullBF.trend

 logbf.int logbf.trend
 0.5841906    1.683028

> exp(logIntervalNullBF.trend)

 logbf.int logbf.trend
  1.793539     5.38183
```

[8]The same process holds for the JZS+AR Bayes factor, but we only show the TAR Bayes factor for brevity.

Name in R	Name in manuscript	Column number in chains
`mu0`	μ_0	1
`sig*delta`	$\sigma_z\delta$	2
`beta0`	β_0	3
`sig*beta1`	$\sigma_z\beta_1$	4
`sig2`	σ_z^2	5
`rho`	ρ	8

Table 2.1: Columns of interest in the `chains.trend` matrix, along with their names in the manuscript and column numbers.

As before, every time the code above is run, the values will be slightly different, due to the random nature of MCMC estimation.

The variable `chains.trend` contains a matrix with eight columns, one for each parameter of the trend model. Each row represents an MCMC sample from the posterior distribution of a parameter. The parameters likely of interest to researchers are shown in Table 2.1 of this supplement. We can draw histograms of the samples for the parameters, as shown in Figure 2.15. These histograms are approximations to the posterior distributions: for example, we can draw a histrogram for the ρ parameter from the trend model:

```
> hist(chains.trend[, 8], main =
        "Posterior for autocorrelation coeff.",
         xlim = c(0, 1))
```

It is also easy to get posterior summary statistics:

```
> summary(chains.trend[, 8])

Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean             SD       Naive SE Time-series SE
      0.135222       0.104743       0.001047       0.002998

2. Quantiles for each variable:

    2.5%      25%      50%      75%    97.5%
0.004436 0.052011 0.114532 0.195980 0.395421
```

In addition to the `ttest.Gibbs.AR()` and `trendtest.Gibbs.AR()` functions, the BayesSingleSub package also contains the `ttest.MCGQ.AR()` and

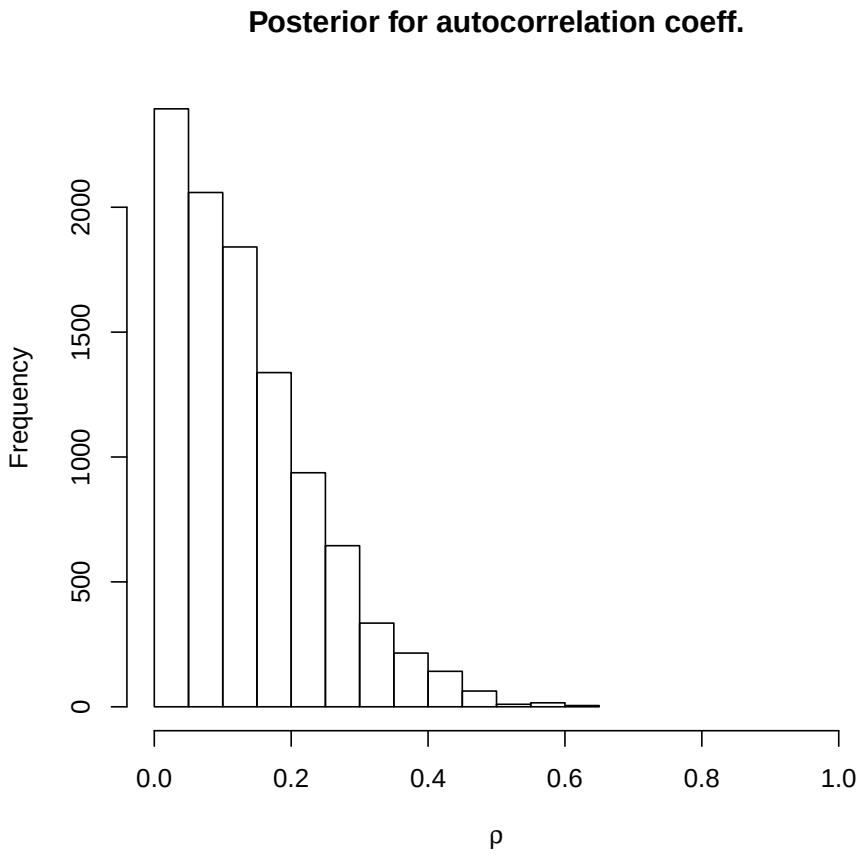

Figure 2.15: Posterior distribution for ρ in the example data from Figure 1 in the manuscript.

`trendtest.MC.AR()` functions. We have not discussed these functions so far because they do not provide qualitatively different information from the information provided by the `ttest.Gibbs.AR()` and `trendtest.Gibbs.AR()` functions, and we did not want to confuse the reader by discussing several functions simultaneously. However, the `ttest.MCGQ.AR()` and `trendtest.MC.AR()` functions provide faster estimates of the Bayes factors than the `ttest.Gibbs.AR()` and `trendtest.Gibbs.AR()` functions, respectively. Their only disadvantage is that they do not return posterior distributions or interval null Bayes factors. This is because they estimate the Bayes factors by using Monte Carlo integration rather than Gibbs sampling. But when only the Bayes factor estimates are required, the `ttest.MCGQ.AR()` and `trendtest.MC.AR()` functions can be used, rather than the slower `ttest.Gibbs.AR()` and `trendtest.Gibbs.AR()` functions. As before, the functions require a definition of the Phase 1 and Phase 2 data and the number of iterations:

```
> logBAR = ttest.MCGQ.AR(ypre, ypost, iterations = 10000,
         r.scale = 1,  betaTheta = 5)
> logBTRENDS = trendtest.MC.AR(ypre, ypost,
         iterations = 10000, r.scaleInt = 1,
         r.scaleSlp = 1,  betaTheta = 5)
```

Again, the resulting log Bayes factors can be exponentiated to obtain the Bayes factors, and the values for r and b can be changed by changing the `r.scale`, `r.scaleInt`, `r.scaleSlp`, and `betaTheta` arguments. For comparison with the previously computed Bayes factors, we print the new Bayes factors:

```
> logBAR

[1] -0.3672968

> logBTRENDS

 logbf.joint logbf.trend logbf.int
    1.673226    1.392421 0.4869526

> exp(logBAR)

[1] 0.692604

> exp(logBTRENDS)

 logbf.joint logbf.trend logbf.int
    5.329332    4.024583  1.627349
```

Although they are somewhat different from the Bayes factors computed using the `ttest.Gibbs.AR()` and `trendtest.Gibbs.AR()` functions due to random sampling, they are similar. Increasing the number of iterations will give more precise results, which will have a greater level of agreement.

2.7.2 Technical details for estimation

In this section, we detail several different ways of computing the Bayes factors for both the JZS+AR model and the TAR model. Each of these algorithms is implemented in the `BayesSingleSub` R package, which may be installed within R. Additionally, the complete source code is freely available from the Comprehensive R Archive Network (CRAN). Unless otherwise stated, the parameters and constants used in this appendix are defined in the main body of the manuscript.

For both the JZS+AR model and the TAR model, we choose two methods for computing the Bayes factor: a Gibbs sampler and Savage-Dickey estimate, and a Monte Carlo estimator. The Monte Carlo estimate is much faster than the Gibbs sampler, but does not provide posterior distributions for the parameters, nor can it be used to compute Bayes factors with interval nulls. However, although the Gibbs sampler is slower, it is still fairly fast; we recommend the Gibbs sampler be used by default. We include the Monte Carlo estimate primarily as a check for the Gibbs sampler estimate.

Computing Bayes factors for the JZS+AR model

Gibbs sampler and the Savage-Dickey density ratio Under certain conditions[9], the Bayes factor for a point-null restriction on a parameter in a more general model can be expressed using the marginal prior and posterior for that parameter. In the case of the JZS+AR model, we seek the Bayes factor for the model with the restriction $\delta = 0$ against the unrestricted model. Given this common situation, the Savage-Dickey identity (Dickey and Lientz, 1970; Morey et al., 2011) relates the Bayes factor to the ratio of the marginal prior distribution to the marginal posterior distribution at the restriction $\delta = 0$ within the unrestricted model; that is,

$$B_{01} = \frac{p_\delta(0 \mid \boldsymbol{y})}{p_\delta(0)}$$

where $p_\delta(0 \mid \boldsymbol{y})$ represents the marginal posterior density function for δ at 0, and $p_\delta(0)$ represents the marginal prior density function for δ at 0.

Figure 2.16 shows this graphically. The dashed line represents the prior distribution for δ in the unrestricted model, which is a Cauchy distribution. The prior density at $\delta = 0$ is $(r\pi)^{-1} = .32$. The solid line shows a hypothetical posterior distribution for δ, again from the unrestricted model. The posterior density at $\delta = 0$ is .11, resulting in a Savage-Dickey density ratio of $.11/.32 = .34$. The Savage-Dickey identity implies that the evidence for the restriction $\delta = 0$, as measured by the Bayes factor, is precisely the degree to which the data changes the density at $\delta = 0$, from prior to posterior. For more details on the use of the Savage-Dickey density ratio, see Wagenmakers et al. (2010) and Morey et al. (2011).

[9]These conditions are technical, but amount to a requirement that the joint prior distribution under the null model is the same as the joint prior distribution under the alternative for all nuisance parameters when the restriction holds. Our stipulation that the prior on δ is independent of the other priors ensures this condition holds.

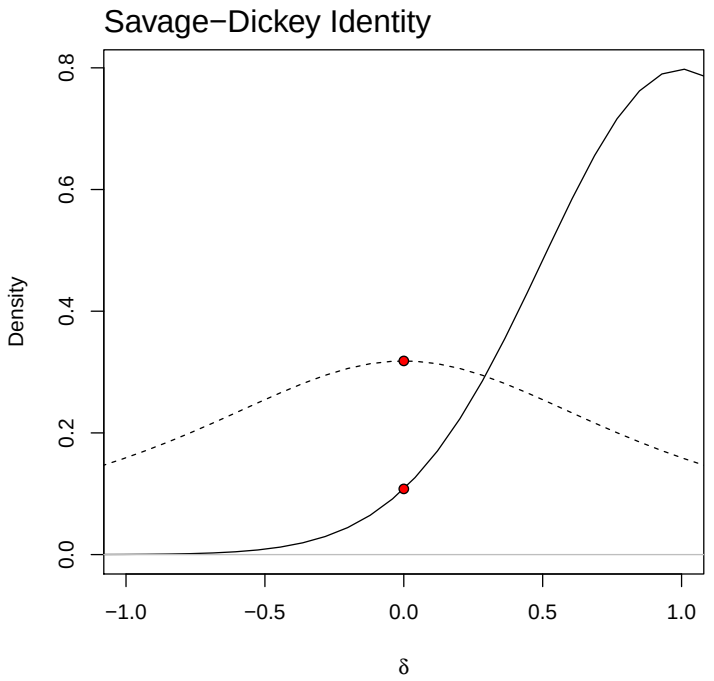

Figure 2.16: Savage-Dickey identity. Dashed line represents the Cauchy prior distribution for δ, solid line represents a hypothetical posterior distribution for δ, both for the unrestricted model. Red dots are the densities at $\delta = 0$. The density at 0 changes by a factor of about 1/3 from prior for posterior, indicating that the Bayes factor for the restriction $\delta = 0$ is 1/3.

In order to estimate the JZS+AR Bayes factor using the Savage-Dickey density ratio, it is necessary to estimate the marginal posterior density for δ, and to estimate its density at $\delta = 0$. To obtain an estimate of the posterior density, we use a Gibbs sampler (Geman and Geman, 1984), and to estimate the density at 0, we use the density estimation method described by Gelfand and Smith (1990).

The Gibbs sampler is a method of obtaining samples from the joint posterior distribution of all parameters. As an example, suppose we have three parameters, θ_1, θ_2, and θ_3. The Gibbs sampler works by assuming exact knowledge — initially, starting values — of all parameters except one. The single unknown parameter, say θ_1, will have a posterior distribution, called a "full conditional posterior" and denoted $p(\theta_1 \mid \theta_2, \theta_3)$, because it is conditioned on knowledge of all other parameters. We sample from this full conditional posterior distribution to obtain a new value for θ_1. We do the same for θ_2 conditioned on θ_3 and our new value of θ_1, and θ_3 conditioned on the new value of θ_1 and θ_2. The samples of θ_1, θ_2, and θ_3 are jointly taken as a sample from the joint posterior. The process is then repeated. On each iteration, we obtain a new sample from the joint posterior of θ_1,θ_2, and θ_3. If certain weak conditions are met, the observed distribution of samples will approach the true joint posterior (Ross, 2002). This is called convergence. Often the samples at the start of the sampling process, called the burn in phase, are discarded from the data set, to ensure that only the samples from the true joint posterior are included. However, when convergence is quick and the number of iterations is large, the effect of these early samples on the final observed distribution is minimal.

We give here the full conditional posterior distributions necessary for building a Gibbs sampler. Due to the conjugacy of the priors[10], proofs are trivial, and we omit them for brevity. We refer interested readers to introductory Bayesian texts such as Gelman et al. (2004). Rouder and Lu (2005) provide a tutorial on the Gibbs sampler for psychologists.

In the full conditionals below, the dot $\cdot$ represents the data and all parameters except the parameter for which the full conditional is defined, and $\mathbf{1}$ represents an $N \times 1$ column vector.

- **Full conditional posterior distribution for** μ_0: Let

$$\begin{aligned} \mu_{\mu_0} &= \frac{(\boldsymbol{y} - \alpha_1 \boldsymbol{x})' \boldsymbol{\Psi}^{-1} \mathbf{1}}{\mathbf{1}' \boldsymbol{\Psi}^{-1} \mathbf{1}}, \text{ and} \\ \sigma^2_{\mu_0} &= \frac{\sigma_z^2}{\mathbf{1}' \boldsymbol{\Psi}^{-1} \mathbf{1}}. \end{aligned}$$

  The full conditional distribution of μ_0 given all other parameters and data $\boldsymbol{y}$ is Normal:

$$\mu_0 \mid \cdot \sim \text{Normal}(\mu_{\mu_0}, \sigma^2_{\mu_0}).$$

[10] Choosing conjugate priors means choosing priors such that the full conditional posteriors are members of the same family of distributions as the priors. This is due to the prior and likelihood having a complementary mathematical form.

- **Full conditional posterior distribution for** α_1: Let

$$\begin{aligned} \mu_{\alpha_1} &= \frac{(\boldsymbol{y} - \mu_0 \mathbf{1})' \boldsymbol{\Psi}^{-1} \boldsymbol{x}}{\boldsymbol{x}' \boldsymbol{\Psi}^{-1} \boldsymbol{x} + \frac{1}{g}}, \text{ and} \\ \sigma^2_{\alpha_1} &= \frac{\sigma^2_z}{\boldsymbol{x}' \boldsymbol{\Psi}^{-1} \boldsymbol{x} + \frac{1}{g}}. \end{aligned}$$

  The full conditional distribution of α_1 given all other parameters and data $\boldsymbol{y}$ is Normal:

$$\alpha_1 \mid \cdot \sim \text{Normal}(\mu_{\alpha_1}, \sigma^2_{\alpha_1}).$$

- **Full conditional posterior distribution for** σ^2_z: Let

$$\begin{aligned} \alpha_{\sigma^2_z} &= \frac{N+1}{2}, \text{ and} \\ \beta_{\sigma^2_z} &= \frac{1}{2}\left(\frac{\alpha_1^2}{g} + (\boldsymbol{y} - \mu_0 \mathbf{1} - \alpha_1 \boldsymbol{x})' \boldsymbol{\Psi}^{-1} (\boldsymbol{y} - \mu_0 \mathbf{1} - \alpha_1 \boldsymbol{x}) \right). \end{aligned}$$

  The full conditional distribution of σ^2_z given all other parameters and data $\boldsymbol{y}$ is Inverse Gamma:

$$\sigma^2_z \mid \cdot \sim \text{Inverse Gamma}(\alpha_{\sigma^2_z}, \beta_{\sigma^2_z}).$$

- **Full conditional posterior distribution for** g: Let

$$\begin{aligned} \alpha_g &= 1, \text{ and} \\ \beta_g &= \frac{\alpha_1^2}{2\sigma^2_z} + \frac{r^2}{2}. \end{aligned}$$

  The full conditional distribution of g given all other parameters and data $\boldsymbol{y}$ is Inverse Gamma:

$$g \mid \cdot \sim \text{Inverse Gamma}(\alpha_g, \beta_g).$$

  The g parameter was only mentioned in passing in the manuscript; it is a convenience parameter, included only to make the sampling easier. The justification for including g is that the Cauchy prior on δ can by constructed using a mixture of Normals. If

$$\begin{aligned} \alpha_1 \mid g, \sigma^2_z &\sim \text{Normal}(0, g\sigma^2_z), \text{and} \\ g &\sim \text{Inverse Gamma}(1/2, r^2/2), \end{aligned}$$

  then

$$\delta = \frac{\alpha_1}{\sigma_z} \sim \text{Cauchy}(r),$$

  as desired.

- **Full conditional posterior distribution for** ρ: The full conditional distribution of ρ given all other parameters and data $\boldsymbol{y}$ does not have a familiar form. Its density function is known up to a proportionality constant:

$$\begin{aligned} & p(\rho \mid \cdot) \propto \\ & (1-\rho)^{b-1} \left|\boldsymbol{\Psi}^{-1}\right|^{\frac{1}{2}} \times \\ & \exp\left\{-\frac{1}{2\sigma_z^2}(\boldsymbol{y}-\mu_0\mathbf{1}-\alpha_1\boldsymbol{x})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\mu_0\mathbf{1}-\alpha_1\boldsymbol{x})\right\}. \end{aligned}$$

  Because the full conditional posterior for ρ does not have a known form, we use the random-walk Metropolis-Hastings algorithm (Ross, 2002) to obtain samples from its full conditional posterior distribution.

The Gibbs sampler proceeds from starting values (which can be estimated from the data) and sampling from each of these full conditionals in turn. The full conditional posterior for the standardized mean difference δ, required for estimating the Savage-Dickey density ratio, can be obtained by dividing the full conditional posterior for α_1 by σ_z:

$$\delta \mid \cdot \quad \sim \quad \text{Normal}(\mu_\delta, \sigma_\delta^2), \tag{2.2}$$

where

$$\begin{aligned} \mu_\delta &= \frac{(\boldsymbol{y}-\mu_0\mathbf{1})'\boldsymbol{\Psi}^{-1}\boldsymbol{x}}{\sigma_z(\boldsymbol{x}'\boldsymbol{\Psi}^{-1}\boldsymbol{x}+\frac{1}{g})}, \text{ and} \\ \sigma_\delta^2 &= \frac{1}{\boldsymbol{x}'\boldsymbol{\Psi}^{-1}\boldsymbol{x}+\frac{1}{g}}. \end{aligned}$$

In order to obtain an estimate of the marginal posterior density of δ at 0, we use the density estimate suggested by Gelfand and Smith (1990). On every iteration of the Gibbs sampler, we compute the full conditional density of δ (Eq. 2.2) at $\delta = 0$. This will yield a chain of density values $d_m, m = 1, \ldots, M$, where M is the total number of Gibbs sampler iterations, possibly after removal of the burn in iterations. We can then estimate the marginal posterior density at $\delta = 0$ by

$$\hat{p_\delta}(0 \mid \boldsymbol{y}) = \frac{1}{M}\sum_{m=1}^{M} d_m.$$

Because the marginal prior density of δ at 0 is known analytically from the density function of the Cauchy distribution, the Savage-Dickey estimate of the Bayes factor is

$$B_{AR} = \frac{\hat{p_\delta}(0 \mid \boldsymbol{y})}{p_\delta(0)}.$$

Monte Carlo integration Another strategy for computing the Bayes factor is to compute the marginal likelihoods of the null and alternative models, and then form the ratio of these marginal likelihoods. In this section, we describe how the two marginal likelihoods can be computed.

The marginal likelihood for the null model, L_0, is

$$L_0 = \int\int\int p_0(\boldsymbol{y} \mid \mu_0, \sigma_z^2, \rho) p(\mu_0) p(\sigma_z^2) p(\rho) d\mu_0 \, d\sigma_z^2 \, d\rho,$$

where p_0 is the likelihood function under the null model, and $p(\mu_0)$, $p(\sigma_z^2)$, and $p(\rho)$ are the prior densities for their respective parameters. Parameters μ_0 and σ_z^2 may be analytically integrated out, leaving a one-dimensional integral over ρ:

$$\begin{aligned} L_0 = & \int_0^1 \pi^{-\frac{N-1}{2}} \left|\boldsymbol{\Psi}^{-1}\right|^{\frac{1}{2}} \left(\mathbf{1}'\boldsymbol{\Psi}^{-1}\mathbf{1}\right)^{-\frac{1}{2}} \Gamma\left(\frac{N-1}{2}\right) \times \\ & \left[\boldsymbol{y}'\left(\boldsymbol{\Psi}^{-1} - \frac{\boldsymbol{\Psi}^{-1}\boldsymbol{J}\boldsymbol{\Psi}^{-1}}{\mathbf{1}'\boldsymbol{\Psi}^{-1}\mathbf{1}}\right)\boldsymbol{y}\right]^{-\frac{N-1}{2}} p(\rho) d\rho, \end{aligned}$$

where $\boldsymbol{J}$ is an $N \times N$ matrix of ones (i.e., $\mathbf{1}\mathbf{1}'$). The derivation of this expression is straightforward, and we omit it for brevity. The integral L_0 can be performed quickly and accurately using a numerical integration technique called Gaussian quadrature (Press et al., 1992).

Let $\alpha_1 = \sigma_z \delta$. The marginal likelihood for the alternative model, L_1, is slightly more complicated:

$$\begin{aligned} L_1 = & \int\cdots\int p_1(\boldsymbol{y} \mid \mu_0, \alpha_1, \sigma_z^2, g, \rho) \times \\ & p(\alpha_1 \mid g, \sigma_z^2) p(\mu_0) p(\sigma_z^2) p(g) p(\rho) d\mu_0 \, d\alpha_1 \, d\sigma_z^2 \, dg \, d\rho, \end{aligned}$$

where p_1 is the likelihood function under the alternative model. Parameters μ_0, α_1, and σ_z^2 may be analytically integrated out, leaving a two-dimensional integral over g and ρ:

$$\begin{aligned} L_1 = & \int_0^1 \int_0^\infty \pi^{-\frac{N-1}{2}} \left|\boldsymbol{\Psi}^{-1}\right|^{\frac{1}{2}} \left(\mathbf{1}'\boldsymbol{\Psi}^{-1}\mathbf{1}\right)^{-\frac{1}{2}} \times \\ & \left(\boldsymbol{x}'\boldsymbol{\Psi}_1^{-1}\boldsymbol{x} + \frac{1}{g}\right)^{-\frac{1}{2}} \Gamma\left(\frac{N-1}{2}\right) \left[\boldsymbol{y}'\boldsymbol{\Psi}_2^{-1}\boldsymbol{y}\right]^{-\frac{N-1}{2}} \times \\ & g^{-\frac{1}{2}} p(g) p(\rho) dg \, d\rho, \end{aligned} \tag{2.3}$$

where

$$\begin{aligned} \boldsymbol{\Psi}_2^{-1} &= \boldsymbol{\Psi}_1^{-1} - \frac{\boldsymbol{\Psi}_1^{-1}\boldsymbol{x}\boldsymbol{x}'\boldsymbol{\Psi}_1^{-1}}{\boldsymbol{x}'\boldsymbol{\Psi}_1^{-1}\boldsymbol{x} + \frac{1}{g}}, \\ \boldsymbol{\Psi}_1^{-1} &= \boldsymbol{\Psi}^{-1} - \frac{\boldsymbol{\Psi}^{-1}\boldsymbol{J}\boldsymbol{\Psi}^{-1}}{\mathbf{1}'\boldsymbol{\Psi}^{-1}\mathbf{1}}. \end{aligned}$$

Because two-dimensional quadrature is substantially more challenging than one-dimensional quadrature, we do not use quadrature to estimate L_1. It is apparent from Eq. 2.3 that L_1 may be interpreted as an expected value with respect to the prior distributions for g and ρ:

$$\begin{aligned} L_1 &= E_{g\rho}\left[f(g,\rho)\right], \qquad (2.4) \\ f_1(g,\rho) &= \pi^{-\frac{N-1}{2}} \left|\boldsymbol{\Psi}^{-1}\right|^{\frac{1}{2}} \left(\mathbf{1}'\boldsymbol{\Psi}^{-1}\mathbf{1}\right)^{-\frac{1}{2}} \left(\boldsymbol{x}'\boldsymbol{\Psi}_1^{-1}\boldsymbol{x} + \frac{1}{g}\right)^{-\frac{1}{2}} \times \\ &\quad \Gamma\left(\frac{N-1}{2}\right) \left[\boldsymbol{y}'\left(\boldsymbol{\Psi}_1^{-1} - \frac{\boldsymbol{\Psi}_1^{-1}\boldsymbol{x}\boldsymbol{x}'\boldsymbol{\Psi}_1^{-1}}{\boldsymbol{x}'\boldsymbol{\Psi}_1^{-1}\boldsymbol{x}} + \frac{1}{g}\right)\boldsymbol{y}\right]^{-\frac{N-1}{2}} \times \\ &\quad g^{-\frac{1}{2}}. \end{aligned}$$

Eq. 2.4 suggests a Monte Carlo estimate of L_1:

$$\hat{L}_1 = \frac{1}{M}\sum_{m=1}^{M} f_1(g_m^*, \rho_m^*), \qquad (2.5)$$

where g_m^* and ρ_m^* are sequences of M independent samples from the prior distributions of g and ρ, respectively. The law of large numbers guarantees that $\hat{L}_1$ will converge to L_1 as $M \to \infty$.

Because the Bayes factor is the ratio of marginal likelihoods L_0/L_1, we can estimate the Bayes factor with $L_0/\hat{L}_1$. This Monte Carlo estimate of the Bayes factor is extremely fast and reliable, but unlike the estimate obtained from the Gibbs sampler, it is not accompanied by parameter estimates and cannot be used to compute Bayes factors for interval null hypotheses.

Computing the Bayes factor for the TAR model

Gibbs sampler and the Savage-Dickey density ratio The Gibbs sampler for the TAR model is substantially the same as for the JZS+AR model, with several exceptions. First, it is helpful to derive multivariate full conditionals when possible, in order to sample them as a block of parameters. This practice, called "blocking", yields more efficient Gibbs sampler chains (Roberts and Sahu, 1997). We use this technique for μ_0, α_1, β_0, and α_2. In addition, we now have two parameters of interest (α_1 and α_2, the unstandardized intercept and slope differences); these two parameters have mutually independent Cauchy priors placed on them. We construct the Cauchy prior as with the JZS+AR model, using mixtures of Normals. We thus need two g parameters, g_1 and g_2, and two scale parameters, r_1 and r_2, used in the priors for α_1 and α_2, respectively.

We first define some necessary terms. Let

$$\begin{aligned}
\boldsymbol{B} &= \begin{pmatrix} \mu_0 \\ \alpha_1 \\ \beta_0 \\ \alpha_2 \end{pmatrix}, \\
\boldsymbol{G}^{-1} &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{g_1} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{g_2} \end{pmatrix}, \text{ and} \\
\boldsymbol{X} &= (\mathbf{1}\ \boldsymbol{x}\ \boldsymbol{t}\ \boldsymbol{\gamma}),
\end{aligned}$$

where $\boldsymbol{\gamma}$ is the column vector whose ith row contains $x_i t_i$. The full conditional posteriors for $\boldsymbol{B}$, σ_z^2, g_1, g_2, and ρ in the TAR model are:

- **Full conditional posterior distribution for $\boldsymbol{B}$**: Let

$$\begin{aligned}
\boldsymbol{\mu_B} &= \left(\boldsymbol{X'\Psi^{-1}X} + \boldsymbol{G^{-1}}\right)^{-1} \boldsymbol{X'\Psi^{-1}y}, \text{ and} \\
\boldsymbol{\Sigma_B} &= \sigma_z^2(\boldsymbol{X'\Psi^{-1}X} + \boldsymbol{G^{-1}})^{-1}.
\end{aligned}$$

  The full conditional distribution of $\boldsymbol{B}$ given all other parameters and data $\boldsymbol{y}$ is Multivariate Normal:

$$\boldsymbol{B} \mid \cdot \ \sim \ \text{MvtNormal}(\boldsymbol{\mu_B}, \boldsymbol{\Sigma_B}).$$

- **Full conditional posterior distribution for σ_z^2**: Let

$$\begin{aligned}
\alpha_{\sigma_z^2} &= \frac{N+2}{2}, \text{ and} \\
\beta_{\sigma_z^2} &= \frac{1}{2}\left((\boldsymbol{y} - \boldsymbol{XB})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{XB}) + \boldsymbol{B'G^{-1}B}\right).
\end{aligned}$$

  The full conditional distribution of σ_z^2 given all other parameters and data $\boldsymbol{y}$ is Inverse Gamma:

$$\sigma_z^2 \mid \cdot \ \sim \ \text{Inverse Gamma}(\alpha_{\sigma_z^2}, \beta_{\sigma_z^2}).$$

- **Full conditional posterior distribution for g_1**: Let

$$\beta_{g_1} = \frac{\alpha_1^2}{2\sigma_z^2} + \frac{r_1^2}{2}.$$

  The full conditional distribution of g_1 given all other parameters and data $\boldsymbol{y}$ is Inverse Gamma:

$$g_1 \mid \cdot \ \sim \ \text{Inverse Gamma}(1, \beta_{g_1}).$$

- **Full conditional posterior distribution for** g_2: Let

$$\beta_{g_2} = \frac{\alpha_2^2}{2\sigma_z^2} + \frac{r_2^2}{2}.$$

The full conditional distribution of g_2 given all other parameters and data $\boldsymbol{y}$ is Normal:

$$g_2 \mid \cdot \sim \text{Inverse Gamma}(1, \beta_{g_2}).$$

- **Full conditional posterior distribution for** ρ: As in the JZS+AR model, the full conditional distribution of ρ given all other parameters and data $\boldsymbol{y}$ does not have a familiar form. Its density function is known up to a proportionality constant:

$$p(\rho \mid \cdot) \propto (1-\rho)^{b-1} \left|\boldsymbol{\Psi}^{-1}\right|^{\frac{1}{2}} \times \exp\left\{-\frac{1}{2\sigma_z^2}(\boldsymbol{y}-\boldsymbol{XB})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{XB})\right\}.$$

We use random-walk Metropolis-Hastings to sample from the full conditional posterior distribution for ρ.

The full conditional posteriors for the standardized effects δ, β_1, and δ and β_1 jointly can be obtained by dividing the full conditional posteriors for α_1, α_2, and α_1 and α_2 jointly, respectively, by σ_z. The full conditional for each of the standardized parameters follows:

- **Full conditional posterior distribution for** δ: Let

$$\mu_\delta = \frac{\boldsymbol{x}'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\mu_0\mathbf{1}-\beta_0\boldsymbol{t}-\alpha_2\boldsymbol{\gamma})}{\sigma_z\left(\boldsymbol{x}'\boldsymbol{\Psi}^{-1}\boldsymbol{x}+\frac{1}{g_1}\right)}, \text{ and}$$

$$\sigma_\delta^2 = \frac{1}{\boldsymbol{x}'\boldsymbol{\Psi}^{-1}\boldsymbol{x}+\frac{1}{g_1}}.$$

The full conditional distribution of δ given all other parameters and data $\boldsymbol{y}$ is Normal:

$$\delta \mid \cdot \sim \text{Normal}(\mu_\delta, \sigma_\delta^2).$$

- **Full conditional posterior distribution for** β_1: Let

$$\mu_{\beta_1} = \frac{\boldsymbol{\gamma}'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\mu_0\mathbf{1}-\alpha_1\boldsymbol{x}-\beta_0\boldsymbol{t})}{\sigma_z\left(\boldsymbol{\gamma}'\boldsymbol{\Psi}^{-1}\boldsymbol{\gamma}+\frac{1}{g_2}\right)}, \text{ and}$$

$$\sigma_{\beta_1}^2 = \frac{1}{\boldsymbol{\gamma}'\boldsymbol{\Psi}^{-1}\boldsymbol{\gamma}+\frac{1}{g_2}}.$$

The full conditional distribution of β_1 given all other parameters and data $\boldsymbol{y}$ is Normal:

$$\beta_1 \mid \cdot \quad \sim \quad \text{Normal}(\mu_{\beta_1}, \sigma^2_{\beta_1}).$$

- **Full conditional posterior distribution for** $(\delta, \beta_1)'$: Let

$$\begin{aligned} \boldsymbol{\mu}_{\delta\beta_1} &= \frac{\left(\boldsymbol{X}_1'\boldsymbol{\Psi}^{-1}\boldsymbol{X}_1 + \boldsymbol{G}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{\Psi}^{-1}\left(\boldsymbol{y} - \mu_0\mathbf{1} - \beta_0\boldsymbol{t}\right)}{\sigma_z}, \\ &\text{and} \\ \boldsymbol{\Sigma}_{\delta\beta_1} &= \left(\boldsymbol{X}_1'\boldsymbol{\Psi}^{-1}\boldsymbol{X}_1 + \boldsymbol{G}_1^{-1}\right)^{-1}. \end{aligned}$$

Also let $\boldsymbol{X}_1$ be a design matrix, created using the two columns of the full design matrix $\boldsymbol{X}$ corresponding to the parameters α_1 and α_2 (that is, the second and fourth columns), and $\boldsymbol{G}_1$ be a 2×2 diagonal matrix whose diagonal is (g_1, g_2). The full conditional distribution of $(\delta, \beta_1)'$ given all other parameters and data $\boldsymbol{y}$ is Multivariate Normal:

$$\begin{pmatrix} \delta \\ \beta_1 \end{pmatrix} \mid \cdot \quad \sim \quad \text{MvtNormal}_2(\boldsymbol{\mu}_{\delta\beta_1}, \boldsymbol{\Sigma}_{\delta\beta_1}).$$

The marginal posterior density for each of the restrictions ($\delta = 0$, $\beta_1 = 0$, and $\delta = \beta_1 = 0$) can be obtained in a manner analogous to the method used for the JZS+AR Bayes factor. The marginal prior distributions for δ and β_1 are mutually independent Cauchy distributions, yielding analytical expressions for the marginal prior density at the various restrictions. The Savage-Dickey estimate of the Bayes factor follows.

Monte Carlo integration As with the JZS+AR Bayes factor in the previous section, we can estimate the Bayes factor using Monte Carlo integration. As previously, the marginal likelihood for the fully null model is a one dimensional integral over ρ, and may be computed using Gaussian quadrature. We call the marginal likelihood for the fully null model L_0[11].

Unlike with the JZS+AR Bayes factor, however, there are three alternative models instead of one. For each of these alternative models the marginal likelihood must be computed. We show here how we estimate the marginal likelihood for the full alternative model, where both differences in intercept and slope are possible. The remaining marginal likelihoods can be estimated analogously.

Again, let $\alpha_1 = \sigma_z\delta$, and let $\alpha_2 = \sigma_z\beta_1$. The marginal likelihood for the full

[11] Although we use the same symbol for the null marginal likelihood in the AR and intercept-slope models, they are different: in the intercept-slope model, the fully null model contains a parameter for the overall trend, whereas in the AR model, this parameter is absent.

alternative model, L_f, is:

$$\begin{aligned} L_f &= \int \cdots \int p_1(\boldsymbol{y} \mid \mu_0, \alpha_1, \beta_0, \alpha_2, \sigma_z^2, g_1, g_2, \rho) \times \\ & \quad p(\alpha_1 \mid g_1, \sigma_z^2) p(\alpha_2 \mid g_2, \sigma_z^2) \times \\ & \quad p(\mu_0) p(\beta_0) p(\sigma_z^2) p(g_1) p(g_2) p(\rho) \times \\ & \quad d\mu_0 \, d\alpha_1 \, d\beta_0 \, d\alpha_2 \, d\sigma_z^2 \, dg_1 \, dg_2 \, d\rho \end{aligned}$$

where p_1 is the likelihood function under the alternative model. Parameters μ_0, β_0, α_1, α_2, and σ_z^2 may be analytically integrated out, leaving a three-dimensional integral over g_1, g_2, and ρ:

$$\begin{aligned} L_f &= \int_0^1 \int_0^\infty \int_0^\infty \pi^{-\frac{N-2}{2}} \left|\boldsymbol{\Psi}^{-1}\right|^{\frac{1}{2}} \left(\mathbf{1}'\boldsymbol{\Psi}^{-1}\mathbf{1}\right)^{-\frac{1}{2}} \times \\ & \quad \left(\boldsymbol{t}'\boldsymbol{\Psi}_1^{-1}\boldsymbol{t}\right)^{-\frac{1}{2}} \left(\boldsymbol{x}'\boldsymbol{\Psi}_2^{-1}\boldsymbol{x} + \frac{1}{g_1}\right)^{-\frac{1}{2}} \left(\boldsymbol{\gamma}'\boldsymbol{\Psi}_3^{-1}\boldsymbol{\gamma} + \frac{1}{g_2}\right)^{-\frac{1}{2}} \times \\ & \quad \Gamma\left(\frac{N-2}{2}\right) \left(\boldsymbol{y}'\boldsymbol{\Psi}_4^{-1}\boldsymbol{y}\right)^{-\frac{N-2}{2}} \times \\ & \quad g_1^{-\frac{1}{2}} g_2^{-\frac{1}{2}} p(g_1) p(g_2) p(\rho) dg_1 \, dg_2 \, d\rho, \end{aligned} \tag{2.6}$$

where

$$\begin{aligned} \boldsymbol{\Psi}_4^{-1} &= \boldsymbol{\Psi}_3^{-1} - \frac{\boldsymbol{\Psi}_3^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{\Psi}_3^{-1}}{\boldsymbol{\gamma}'\boldsymbol{\Psi}_3^{-1}\boldsymbol{\gamma} + \frac{1}{g_2}}, \\ \boldsymbol{\Psi}_3^{-1} &= \boldsymbol{\Psi}_2^{-1} - \frac{\boldsymbol{\Psi}_2^{-1}\boldsymbol{x}\boldsymbol{x}'\boldsymbol{\Psi}_2^{-1}}{\boldsymbol{x}'\boldsymbol{\Psi}_2^{-1}\boldsymbol{x} + \frac{1}{g_1}}, \\ \boldsymbol{\Psi}_2^{-1} &= \boldsymbol{\Psi}_1^{-1} - \frac{\boldsymbol{\Psi}_1^{-1}\boldsymbol{t}\boldsymbol{t}'\boldsymbol{\Psi}_1^{-1}}{\boldsymbol{t}'\boldsymbol{\Psi}_1^{-1}\boldsymbol{t}}, \\ \boldsymbol{\Psi}_1^{-1} &= \boldsymbol{\Psi}^{-1} - \frac{\boldsymbol{\Psi}^{-1}\boldsymbol{J}\boldsymbol{\Psi}^{-1}}{\mathbf{1}'\boldsymbol{\Psi}^{-1}\mathbf{1}}. \end{aligned}$$

Eq. 2.6 represents an expected value with respect to g_1, g_2, and ρ, suggesting the Monte Carlo estimate

$$\hat{L}_f = \sum_{m=1}^{M} f_2(g_{1m}^*, g_{2m}^*, \rho_m^*),$$

where g_{1m}^*, g_{2m}^*, and ρ_m^* are sequences of M independent samples from the prior distributions of g_1, g_2, and ρ, respectively, and

$$\begin{aligned} f_2(g_1, g_2, \rho) &= \pi^{-\frac{N-2}{2}} \left|\boldsymbol{\Psi}^{-1}\right|^{\frac{1}{2}} \left(\mathbf{1}'\boldsymbol{\Psi}^{-1}\mathbf{1}\right)^{-\frac{1}{2}} \left(\boldsymbol{t}'\boldsymbol{\Psi}_1^{-1}\boldsymbol{t}\right)^{-\frac{1}{2}} \times \\ & \quad \left(\boldsymbol{x}'\boldsymbol{\Psi}_2^{-1}\boldsymbol{x} + \frac{1}{g_1}\right)^{-\frac{1}{2}} \left(\boldsymbol{\gamma}'\boldsymbol{\Psi}_3^{-1}\boldsymbol{\gamma} + \frac{1}{g_2}\right)^{-\frac{1}{2}} \times \\ & \quad \Gamma\left(\frac{N-2}{2}\right) \left(\boldsymbol{y}'\boldsymbol{\Psi}_4^{-1}\boldsymbol{y}\right)^{-\frac{N-2}{2}} g_1^{-\frac{1}{2}} g_2^{-\frac{1}{2}}. \end{aligned}$$

Having obtained L_0 via Gaussian quadrature and $\hat{L}_f$ via our Monte Carlo estimate, we can estimate the Bayes factor of the fully null model against the full alternative model with $L_0/\hat{L}_f$. The other Bayes factors can be obtained analogously.